Информатика и её применения

Том 6 Выпуск 2 Год 2012

СОДЕРЖАНИЕ

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2012. Т. 6. Вып. 2. С. 2–13

АЛГОРИТМ ВЫЧИСЛЕНИЯ ХАРАКТЕРИСТИК МОДЕЛИ ТЕЛЕКОММУНИКАЦИОННОЙ СЕТИ С ПОВТОРАМИ ПЕРЕДАЧ И НЕПОЛНОДОСТУПНОЙ СХЕМОЙ УПРАВЛЕНИЯ БУФЕРАМИ*

Я. М. Агаларов[1]

Аннотация: Рассмотрена телекоммуникационная сеть со схемой управления буферами узлов SMQMA (Sharing with Maximum Queue Length and Minimum Allocation) и возможностью повтора передач из источника и в транзитных узлах. Предложен алгоритм расчета усредненных характеристик сети (вероятностей блокировок узлов, суммарной нагрузки на линиях, среднего числа пакетов в узлах и в сети, среднего числа пакетов, ожидающих повтора в источниках, и др.). Приведены доказательства утверждений о свойствах алгоритма и результаты вычислительных экспериментов.

Ключевые слова: телекоммуникационная сеть; повторы передач; механизмы управления буферами

1 Введение

Эффективность принимаемых на этапе проектирования решений по выбору варианта построения телекоммуникационной сети в значительной степени зависит от адекватности используемых моделей процессов функционирования сетей и точности методов их расчета. Для обеспечения адекватности в модели должны быть учтены все существенные (с точки зрения цели исследования) свойства реальной сети. К таким свойствам, в частности, относятся: ограниченность буферной памяти узлов коммутации, механизмы управления буферами и многоканальность линий связи.

Ограниченность буферной памяти узлов коммутации является одной из причин роста числа повторных передач в сетях с коммутацией пакетов и, как следствие, резкого роста нагрузки на отдельные участки сети или сеть в целом и снижение ее производительности.

Значения показателей производительности сети существенным образом зависят и от механизмов управления буферами узлов коммутации, из множества которых наиболее применяемыми на практике являются статические схемы управления [1, 2]: CS (Complete Sharing), CP (Complete Partitioning), SMQ (Sharing with Maximum Queue Length), SMA (Sharing with Minimum Allocation), SMQMA. Схема SMQMA является обобщением первых четырех схем. Для оценки эффективности указанных схем управления буферами разработаны различные модели узлов коммутации (см., например, [1, 3, 4]) и методы расчета их характеристик.

Необходимость учета многоканальности линий связи вызвана тем, что замена в модели многоканальных линий одноканальными может внести значительную погрешность в результаты расчета показателей производительности сети.

В качестве модели сети с ограниченными буферами, как правило, используется сеть массового обслуживания, узлы которой рассматриваются как изолированные системы массового обслуживания (СМО) с ограниченными накопителями и пуассоновскими входящими потоками. Одной из первых работ, где предлагалась подобная модель, является [5], в которой исследовалась задача выбора объемов буферной памяти узлов коммутации сети с датаграммным режимом работы в рамках модели сети с одноканальными линиями связи и схемой распределения буферов CS. В этой же работе для расчета модели был использован подход, который в дальнейшем нашел применение в работах других исследователей (см., например, [6–10]). Суть подхода заключается в следующем. Для стационарных характеристик модели выводится система нелинейных уравнений вида $y_i = f_i(\overline{y}, \overline{a})$, $i = 1, \ldots, n$, где $\overline{y} = (y_1, \ldots, y_n)$ — вектор неизвестных переменных (например, вероятностей блокировки); $\overline{a}$ — вектор известных параметров; f_i — обозначение функции. Для решения указанных уравнений применялись алгоритмы, основанные, как правило, на градиентном методе [4, 5] и методе простой итерации [3, 5–10], вопрос сходимости которых, за исключением редких случаев, остается открытым. Среди работ, в которых для сетей с повторами передач приводится

*Работа выполнена при поддержке РФФИ, грант 11-07-00112.

[1]Институт проблем информатики Российской академии наук, agglar@ya.ru

данный подход и доказательство сходимости алгоритма, следует выделить публикации [4, 6–9], из которых в [6, 8] рассматривалась сеть коммутации каналов, в [4] — случай сети с одноканальными линиями связи, повторами передач из источника и схемой CS, в [7, 9] — случаи сети с многоканальными линиями связи, повторами передач из источника и схемами CS и SMQ соответственно.

Ниже предлагается алгоритм расчета усредненных характеристик (вероятностей блокировки узлов, суммарной нагрузки линий, среднего числа пакетов в узлах и в сети, среднего числа повторов пакета из источника и т. д.) для модели сети со схемой управления SMQMA, без потерь и возможностью повтора пакетов в транзитных узлах и в источниках. Алгоритм основан на указанном выше подходе. Приведено доказательство сходимости алгоритма и численный пример использования алгоритма.

2 Модель сети и постановка задачи

Модель сети представляется в виде графа, состоящего из U вершин и V дуг. Вершины графа отождествляются с узлами связи, дуги — с линиями связи. Заданы множества узлов-входов (узлов коммутации, в которые извне сети поступают пакеты) и узлов-выходов (узлов коммутации, через которые пакеты покидают сеть) и множество нециклических путей L, соединяющих узлы-входы с узлами-выходами, причем для каждой пары узел-вход и узел-выход существует единственный соединяющий их путь. Для каждой пары узел-вход и узел-выход заданы интенсивности внешних потоков пакетов, передача которых в сети происходит по пути, соединяющему эти узлы. Узлы сети имеют ограниченную буферную память со схемой распределения SMQMA, линии связи имеют заданное число однородных каналов.

В качестве модели коммутационного узла используется СМО с ограниченным накопителем (буферной памятью) и несколькими линиями из однотипных каналов, формализованная структура которой приведена на рис. 1.

Отметим, что накопитель разбит на блоки 1, 2 и 3 условно и пакет реально не передается из блока в блок, а хранится в одном месте, прикрепляемом условно к одному из блоков в зависимости от состояния процесса передачи в последующий узел.

Обозначим через v линию связи, u — узел связи, Ω_u^+ — множество исходящих из узла u линий, по которым в последующие узлы передаются пакеты.

В модели выполняются следующие условия:

1. Места в накопителе распределяются согласно схеме SMQMA:
   - за каждой линией v закрепляется $a_v \geq 1$ мест, $\sum\limits_{v\in\Omega_u^+} a_v \leq R_{0u}$;
   - оставшиеся $R_u = R_{0u} - \sum\limits_{v\in\Omega_u^+} a_v$ мест общедоступны;
   - максимальное число общедоступных мест, которые могут занять v-пакеты (пакеты из потока, поступающего на линию v), не должно превышать заданную величину $0 \leq \leq r_v \leq R_u, v \in \Omega_u^+, \sum\limits_{v\in\Omega_u^+} r_v \geq R_u$.
2. Поступившему в СМО пакету предоставляется место в блоке 1 (см. рис. 1), если он передан без ошибок и в момент его поступления в накопителе есть доступное свободное место (произошла успешная попытка передачи из предыдущего узла), иначе пакет получает отказ (произошла неуспешная попытка передачи) и в предыдущем узле может быть повторена попытка его передачи.
3. Количество попыток передачи пакета в узле регулируется вероятностью повтора передачи, а именно: задается вероятность $(1 - a_v)$ того, что попытка передачи является последней для данного v-пакета. Если попытка является не последней (происходит с вероятностью a_v), то в узле хранится копия v-пакета (место хранения которой закрепляется за блоком 1 до отправления и передается блоку 2 или 3 после отправления в последующий узел) и от последующего узла требуется подтверждение успешной передачи, иначе после завершения передачи копия не хранится (пакет освобождает место в накопителе) и подтверждения не требуется. Если пакет принят в накопитель и предыдущий узел требует подтверждения, то в предыдущий узел отправляется подтверждение, где при его получении освобождается занятое пакетом место в накопителе (в блоке 3 предыдущего узла). Если в течение заданного интервала времени (таймаута) с момента отправления пакета (в течение этого времени копия хранится в блоке 2, если попытка не последняя) узел не получает положительного подтверждения, то пакет делает повторную попытку передачи из узла (копия пакета из блока 2 передается в блок 1), иначе делает повторную попытку через случайный интервал времени из источника.

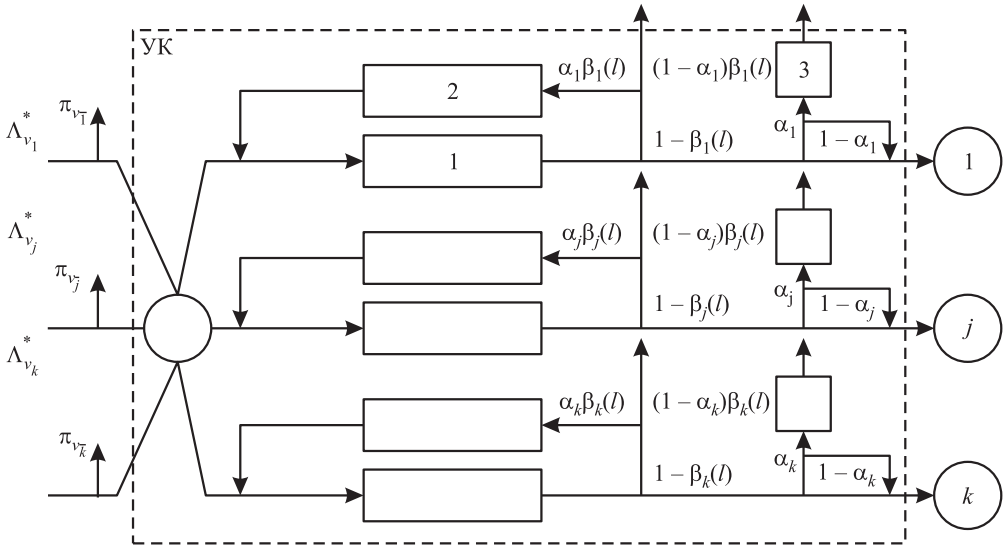


Рис. 1 Модель узла связи: $j = 1, \ldots, k$ — номера линий связи; стрелка — возможное направление «движения» пакета; надпись на стрелке — вероятность, с которой пакет выбирает указанное стрелкой направление; окружность с номером j — сток j-й линии; прямоугольник с номером 1 (блок 1) — место в накопителе узла, где хранятся пакеты (копии пакетов), стоящие в очереди к линии для отправления в последующий узел; прямоугольник с номером 2 (блок 2) — место в накопителе узла, где хранятся неуспешно переданные пакеты, ожидающие повторной передачи; прямоугольник с номером 3 (блок 3) — место в накопителе, где хранятся успешно переданные пакеты, ожидающие подтверждения

4. Если в накопителе освобождается место, закрепленное за линией v, и хотя бы одно общедоступное место в накопителе занимает v-пакет, то освободившееся место становится общедоступным, а одно занятое v-пакетом общедоступное место закрепляется за линией v.
5. Суммарные потоки v-пакетов, поступающих на линии извне узла, являются независимыми в совокупности пуассоновскими потоками. Для обслуживания v-пакета требуется одновременно одно место хранения и один канал типа v, $v \in \Omega_u^+$.
6. Принятые в СМО v-пакеты обслуживаются в порядке поступления.
7. Время занятия канала v-пакетом — экспоненциально распределенная случайная величина с параметром $0 < \mu_v < \infty$, $v \in \Omega_u^+$.
8. Интервал времени $\tau_v > 0$ (тайм-аут), через который неуспешно переданный v-пакет может быть передан повторно, — заданная детерминированная величина.
9. Успешно переданный по линии v пакет освобождает место в накопителе через случайное время t_v с заданным средним значением $\overline{t}_v$, $\tau_v \geq t_v > 0$.
10. Источники пакетов (абонентские узлы) имеют неограниченный накопитель. Передача пакетов по абонентской линии, соединяющей абонентский узел с узлом-входом (узлом-выходом), происходит без ошибок.
11. Подтверждения успешной передачи не теряются.

Введем следующие обозначения:

v^+ — узел-сток линии v;

v^- — узел-исток линии v;

c_v — канальная емкость линии v;

l — путь, соединяющий узел-вход с узлом-выходом;

L_v — множество путей, содержащих линию v, $L_v \subseteq L$;

$v_0(l), \ldots, v_{S_l}(l)$ — линии, составляющие путь l, (т. е. $l = \{v_0, \ldots, v_{S_l}\}$), где $(S_l + 1)$ — число линий на пути l; индексы $0, \ldots, S_l$ показывают порядок следования элементов на пути; v_i — линия, исходящая из узла u_i пути l; $V_0(l)$ — линия, соединяющая источник с узлом-входом; v_{S_l} — абонентская линия, исходящая из узла-выхода u_{S_l};

l_v^+ — множество различных линий, включающее v и линии, следующие после v по направлению к адресату на пути l, включая v_{S_l};

V_v^+ — множество различных линий, включающее $\Omega_{v^-}^+$ и линии $v' \in l_v^+$, $v \in \Omega_{v^-}^+$;

$\lambda(l)$ — интенсивность потока (l-потока) пакетов, поступающих из источника на узел-вход и требующих передачи по пути l, $\lambda(l) > 0$, $l \in L$;

δ_v — вероятность безошибочной передачи пакета по каналу линии v;

$\Lambda^0_{v_{i(l)}}$ — интенсивность суммарного l-потока пакетов на выходе линии $v_{i-1}(l)$, $i = 1, \ldots, S_l$;

$\Lambda^*_v(l)$ — интенсивность суммарного l-потока пакетов, поступающих на линию v;

π_v — вероятность блокировки узла для пакетов, требующих передачи по исходящей линии v.

Пусть $\overline{k}_u = \{\overline{k}_v, v \in \Omega_u^+\}$ — состояние накопителя узла $u \in U$, где $\overline{k}_v = (k_v, k'_v, k''_v)$, k_v — число пакетов в накопителе узла (в блоке 1), стоящих в очереди к линии v, k'_v — число пакетов в накопителе узла (в блоке 2), неуспешно переданных по линии v и ожидающих повторной передачи, k''_v — число пакетов в накопителе узла (в блоке 3), успешно переданных по линии v, ожидающих подтверждения:

$$A_{R_u,\overline{r}_u,\overline{a}_u} = \left\{ \overline{k}_u \colon \sum_{v\in\Omega_u^+} (k_v + k'_v + k''_v - a_v)^+ \le R_u \,, \right.$$
$$\left. k_v + k'_v + k''_v \le r_v + a_v \,, \;\; v \in \Omega_u^+ \right\} ;$$

$$A^{0,m_v}_{R_u,\overline{r}_u,\overline{a}_u} = \left\{ \overline{k}_u \in A_{R_u,\overline{r}_u,\overline{a}_u} : \; k_v + k'_v + k''_v = m_v \right\} ;$$
$$A^{1,m_v}_{R_u,\overline{r}_u,\overline{a}_u} = \left\{ \overline{k}_u \in A_{R_u,\overline{r}_u,\overline{a}_u} : \; k_v = m_v \right\} ;$$
$$A^{2,m_v}_{R_u,\overline{r}_u,\overline{a}_u} = \left\{ \overline{k}_u \in A_{R_u,\overline{r}_u,\overline{a}_u} : \; k'_v + k''_v = m_v \right\} ,$$

где

$$(b)^+ = \begin{cases} b \,, & \text{если } b \ge 0 \,, \\ 0 \,, & \text{если } b < 0 \,, \end{cases}$$
$$m_v = 0, \ldots, r_v + a_v \,.$$

Для данной модели узла с учетом введенных выше обозначений для π_v — вероятности блокировки узла для пакетов, поступающих на линию v, $v \in \Omega_u^+$, в стационарном режиме работы справедлива формула [2, 3]:

$$\pi_v = \frac{1}{g\left(R_u, \overline{a}_u, \overline{r}_u, \overline{\rho}^*_u\right)} \sum_{\overline{k}\in A_{R_u,\overline{r}_u,\overline{a}_u}} p\left(\overline{c}_u, \overline{\pi}_u, \overline{k}_u, \overline{\rho}^*_u\right) . \quad (1)$$

Здесь

$$u = v^- ;$$

$$p\left(\overline{c}_u, \overline{\pi}_u, \overline{k}_u, \overline{\rho}^*_u\right) = \prod_{v\in\Omega_u^+} z_v\left(c_v, \overline{\pi}_{v^+}, \overline{\rho}^*_v, k_v, s_v\right) ,$$

где

$$s_v = k'_v + k''_v \,,$$

$$z_v\left(c_v, \overline{\pi}_{v^+}, \overline{\rho}^*_v, k_v, s_v\right) =$$
$$= \begin{cases} \dfrac{\left(\rho'^*_v + \rho''^*_v\right)^{s_v}}{s_v!} \dfrac{\rho^{*k_v}_v}{k_v!} & \text{при } k_v < c_v \,, \\ \dfrac{\left(\rho'^*_v + \rho''^*_v\right)^{s_v}}{s_v!} \dfrac{\rho^{*k_v}_v}{c_v! c_v^{k_v - c_v}} & \text{при } k_v \ge c_v \,; \end{cases}$$

$$g\left(R_u, \overline{a}_u, \overline{r}_u, \overline{\rho}^*_u\right) = \sum_{\overline{k}_u\in A_{R_u,\overline{r}_u,\overline{a}_u}} p\left(\overline{c}_u, \overline{k}_u, \overline{\rho}^*_u\right) ;$$

$$\overline{\rho}^*_u = \left\{\overline{\rho}^*_v, \; v \in \Omega_u^+\right\} ,$$

где

$$\overline{\rho}^*_v = \left(\rho^*_v, \rho'^*_v, \rho''^*_v\right) ,$$
$$\rho^*_v = \sum_{l\in L_v} \frac{\Lambda^*_v(l)}{\mu_v(1 - \alpha_v\beta_v(l))} \,,$$
$$\rho'^*_v = \sum_{l\in L_v} \frac{\Lambda^*_v(l)\tau_v\alpha_v\beta_v(l)}{1 - \alpha_v\beta_v(l)} \,,$$
$$\rho''^*_v = \sum_{l\in L_v} \frac{\Lambda^*_v(l)\overline{t}_v\alpha_v(1 - \beta_v(l))}{1 - \alpha_v\beta_v(l)} \,,$$

$$\beta_v(l) = 1 - (1 - \delta_v)\left(1 - \pi_{v(l)^+}\right) , \; v \in \Omega_u^+ ;$$

$$\overline{\pi}_u = \left\{\overline{\pi}_{v^+}, \; v \in \Omega_u^+\right\} ,$$

где

$$\overline{\pi}_{v^+} = \left\{\pi_{v'}, v' \in \Omega^+_{v^+}\right\} .$$

Ставится задача расчета значений параметров $\overline{\rho}^*_v, \pi_v, v \in V$.

3 Метод решения

В установившемся режиме работы рассматриваемой сети для потоков в узлах справедливы следующие соотношения:

$$\Lambda^0_{v_j(l)}(l) = \frac{\Lambda^0_{v_{j+1}(l)}(l)}{1 - \beta_{v_{j-1}(l)}(l)} \left(1 - \alpha_{v_j(l)}\beta_{v_j(l)}(l)\right) ,$$
$$j = 1, \ldots, S_l \,, \quad (2)$$

где $\Lambda^0_{v_{S_l+1}(l)}(l) = \lambda(l)$, $\delta_{v_0(l)} = 0$, $\pi_{v_0(l)} = 0$.

Физический смысл формулы заключается в том, что пакет, принятый в накопитель узла $v^-_j(l)$, делает повторные попытки передачи в последующий узел $v^+_j(l)$ в среднем $(1 - \alpha_{v_j(l)}\beta_{v_j(l)}(l))^{-1}$ раз и интенсивность потока таких попыток, делаемых всеми пакетами l-потока, равна величине $\Lambda^*_{v_j(l)}(l)(1 - \beta_{v_{j-1}(l)}(l))/\left(1 - \alpha_{v_j(l)}\beta_{v_j(l)}(l)\right)$.

Из (2) получаем формулы:

$$\Lambda^*_{v_j(l)}(l) = \\ = \frac{\lambda(l)}{1-\pi_{v_i(l)}} \prod_{j=1}^{S_l} \left(\frac{1-\alpha_{v_j(l)}}{1-\beta_{v_j(l)}(l)} + \alpha_{v_j(l)} \right),\ l \in L. \quad (3)$$

Переобозначив в (1) для краткости изложения $1-\pi_v$ через y_v, $z_v\left(c_v, \overline{\pi}_{v^+}, \rho^*_v, k_v, l_v\right)$ через $z_v\left(k_v, l_v, \overline{\Lambda}^*_v, \overline{y}_{v^+}\right)$, $p\left(\overline{c}_u, \overline{\pi}_u, \overline{k}_u, \overline{\rho}^*_u\right)$ через $p_{\overline{k}_u}\left(\overline{\Lambda}^*_u, \overline{y}_u\right)$, $g\left(R_u, \overline{a}_u, \overline{r}_u, \overline{\rho}^*_u\right)$ через $g\left(\overline{a}_u, \overline{\Lambda}^*_u, \overline{y}_u\right)$, выражение в правой части равенства для π_v через $1-q_{R_u,\overline{r}_u,\overline{a}_u}(\overline{\Lambda}^*_u, \overline{y}_u)$, где $u = v^-$, $\overline{\Lambda}^*_u = \left\{\overline{\Lambda}^*_v, v \in \Omega^+_u\right\}$, $\overline{\Lambda}^*_v = \{\Lambda^*_v(l), l \in L_v\}$, $\overline{y}_u = \{\overline{y}_v, v \in \Omega^+_u\}$, $\overline{y}_v = \\ = \{y_{v'}, v' \in \Omega^+_u\}$, получим систему нелинейных уравнений относительно неизвестных переменных y_v:

$$y_v = q_v\left(\overline{\Lambda}^*_u, \overline{y}_u\right),\ v \in V,\ u = v^-. \quad (4)$$

Заметим, что $q_v\left(\overline{\Lambda}^*_u, \overline{y}_u\right)$ — функция, зависящая от $y_{v'}$, $v' \in V^+_v$.

Обозначим набор $\{y_v,\ v \in V\}$ через $\overline{y}$. Будем говорить, что набор $\overline{y}$ положителен, если $y_v \in (0,\ 1]$ для всех $v \in V$.

Пусть задана последовательность $\overline{y}[n] = \\ = \{y_v[n],\ v \in V\}$, $n \geq 0$, где $y_v[n+1] = \\ = q_v(\overline{\Lambda}^*_u[n], \overline{y}_u[n])$, $y_v[0] = 1$, $v \in V$, а $\overline{\Lambda}^*_u[n]$ — это $\overline{\Lambda}^*_u$, вычисленный при $y_{v'} = 1-\pi_{v'} = y_{v'}[n]$, $v' \in l^+_v$, $l \in L_v$. В дальнейшем будем писать $\overline{y}[n+1] < \overline{y}[n]$, если при заданном $n \geq 0$ выполняется неравенство $y_v[n+1] < y_v[n]$ для всех $v \in V$.

Введем обозначения: $\overline{a}_u = \{a_v, v \in \Omega^+_u\}$, $\overline{r}_u = \\ = \{r_v, v \in \Omega^+_u\}$, $\overline{a}_u - \overline{1}_{v'} = \overline{a}'_u = \{a'_v, v \in \Omega^+_u\}$, где

$$a'_v = \begin{cases} a_v & \text{при } v \neq v'; \\ a_v - 1 & \text{при } v = v'. \end{cases}$$

В дальнейшем параметры $(\overline{a}_u, \overline{r}_u)$ будем называть ограничениями доступа узла (СМО).

Теорема. *Существует* $\lim\limits_{n\to\infty} \overline{y}[n] = \overline{y}^0 \geq 0$. *Система* (4) *имеет положительное решение тогда и только тогда, когда* $\lim\limits_{n\to\infty} \overline{y}[n] = \overline{y}^0 > 0$.

Доказательство. Докажем две вспомогательные леммы.

Введем обозначения: $d_{v,\overline{a}_u}\left(\overline{a}_u, \overline{\Lambda}^*_u, \overline{y}\right)$ — среднее число v-пакетов в узле $u = v^-$ с параметрами R_u, $\overline{a}_u$ и $\overline{r}_u$; $d_{1,v,\overline{a}_u}\left(\overline{\Lambda}^*_u, \overline{y}_u\right)$ — среднее число v-пакетов в блоке 1 узла $u = v^-$ с параметрами R_u, $\overline{a}_u$ и $\overline{r}_u$; $d_{2,v,\overline{a}_u}\left(\overline{\Lambda}^*_u, \overline{y}_u\right)$ — среднее число v-пакетов в блоках 2 и 3 узла $u = v^-$ с параметрами R_u, $\overline{a}_u$ и $\overline{r}_u$, $v \in \Omega^+_u$.

Из (1) следуют равенства:

$$d_{v,\overline{a}_u}\left(\overline{a}_u, \overline{\Lambda}^*_u, \overline{y}_u\right) = \\ = \frac{1}{g(\overline{a}_u, \overline{\Lambda}^*_u, \overline{y}_u)} \sum_{m_v=1}^{r_v+a_v} m_v \sum_{\overline{k} \in A^{0,m_v}_{R_u,\overline{r}_u,\overline{a}_u}} p_{\overline{k}_u}(\overline{\Lambda}^*_u, \overline{y}_u); \quad (5)$$

$$d_{1,v,\overline{a}_u}\left(\overline{a}_u, \overline{\Lambda}^*_u, \overline{y}_u\right) = \\ = \frac{1}{g(\overline{a}_u, \overline{\Lambda}^*_u, \overline{y}_u)} \sum_{m_v=1}^{r_v+a_v} m_v \sum_{\overline{k} \in A^{1,m_v}_{R_u,\overline{r}_u,\overline{a}_u}} p_{\overline{k}_u}(\overline{\Lambda}^*_u, \overline{y}_u);$$

$$d_{2,v,\overline{a}_u}\left(\overline{a}_u, \overline{\Lambda}^*_u, \overline{y}_u\right) = \\ = \frac{1}{g(\overline{a}_u, \overline{\Lambda}^*_u, \overline{y}_u)} \sum_{m_v=1}^{r_v+a_v} m_v \sum_{\overline{k} \in A^{2,m_v}_{R_u,\overline{r}_u,\overline{a}_u}} p_{\overline{k}_u}(\overline{\Lambda}^*_u, \overline{y}_u),\ v \in \Omega^+_u.$$

Лемма 1. *В рамках модели, задаваемой соотношениями* (1), *для любого узла* u *и любой линии* $v \in \Omega^+_u$ *справедливы неравенства:*

$$d_{1,v',\overline{a}_u}\left(\overline{a}_u, \overline{\Lambda}^*_u, \overline{y}_u\right) - d_{1,v',\overline{a}_u-1_v}\left(\overline{a}_u, \overline{\Lambda}^*_u, \overline{y}_u\right) > 0;$$
$$d_{2,v',\overline{a}_u}\left(\overline{a}_u, \overline{\Lambda}^*_u, \overline{y}_u\right) - d_{2,v',\overline{a}_u-1_v}\left(\overline{a}_u, \overline{\Lambda}^*_u, \overline{y}_u\right) > 0,$$
$$v' \in \Omega^+_u.$$

Доказательство. Пусть заданы R_u — емкость накопителя; $(\overline{a}_u - 1_v,\ \overline{r}_u)$ — параметры ограничения доступа узла u. Назовем систему с ограничениями доступа $(\overline{a}_u - 1_v,\ \overline{r}_u)$ первой СМО, с ограничениями $(\overline{a}_u, \overline{r}_u)$ — второй СМО и обозначим их соответственно через СМО_1 и СМО_2. Выделим в СМО_2 произвольное место хранения и обозначим его номером $\sum\limits_{v\in\Omega^+_u} a_v + R_u$. Остальные места пронумеруем числами $1, \ldots, \sum\limits_{v\in\Omega^+_u} a_v + R_u - 1$, и пусть пакету присваивается номер соответствующего места, где он хранится. Место с номером $\sum\limits_{v\in\Omega^+_u} a_v + R_u$ могут занимать только v-пакеты.

Рассмотрим СМО_3, отличающуюся от СМО_2 только дисциплиной обслуживания пакетов, и назовем ее третьей СМО. В СМО_3 v-пакеты в первую очередь занимают место с номером $\sum\limits_{v\in\Omega^+_u} a_v + R_u$, если оно свободно, иначе занимают свободные места с номерами $1, \ldots, \sum\limits_{v\in\Omega^+_u} a_v + R_u - 1$. Номер пакета за время пребывания в системе не меняется. Пакеты с номерами $1, \ldots, \sum\limits_{v\in\Omega^+_u} a_v + R_u - 1$ обслуживаются

в каждом из блоков 1, 2, 3 в порядке поступления, и v-пакеты с этими номерами имеют абсолютный приоритет с дообслуживанием перед v-пакетом с номером $\sum\limits_{v\in\Omega_u^+} a_v + R_u$. Заметим, что v-пакет с номером $\sum\limits_{v\in\Omega_u^+} a_v + R_u$ поступает на обслуживание в соответствующий блок только при отсутствии в очереди к этому блоку v-пакетов с меньшими номерами (в блоках 2 и 3 очереди всегда отсутствуют).

Очевидно, интенсивность потока пакетов каждого типа, допускаемых в накопитель СМО_3, больше, чем в СМО_1 (так как занятие v-пакетами места с номером $\sum\limits_{v\in\Omega_u^+} a_v + R_u$ разгружает места с меньшими номерами и не влияет на процессы обслуживания пакетов с этими номерами). А поскольку времена обслуживания в блоках и вероятность повтора пакета каждого типа в СМО_3 и СМО_1 одинаковы, то интенсивность выходного потока каждого из блоков 1, 2, 3 в СМО_3 больше, чем интенсивность выходного потока соответствующего блока СМО_1. Следовательно, среднее число пакетов каждого типа в каждом из блоков 1, 2, 3 в СМО_3 больше, чем в соответствующем блоке СМО_1.

Отметим следующие свойства СМО_3:

- времена обслуживания пакетов каждого типа с номерами из интервала $1, \ldots, \sum\limits_{v\in\Omega_u^+} a_v + R_u - 1$ не зависят от их номеров;
- время дообслуживания v-пакета с номером $\sum\limits_{v\in\Omega_u^+} a_v + R_u$ и время обслуживания любого вновь поступившего v-пакета — одинаково распределенные случайные величины.

Из указанных свойств следует, что если в СМО_3 в момент поступления нового v-пакета менять место его хранения и место хранения обслуживаемого или ожидающего обслуживания в этот момент неприоритетного v-пакета, то процессы изменения числа пакетов каждого типа в системе не изменятся. Таким образом, процессы изменения числа пакетов каждого типа в СМО_3 и СМО_2 совпадают. Следовательно, среднее число пакетов каждого типа в каждом из блоков 1, 2, 3 в СМО_2 больше, чем в соответствующем блоке СМО_1.

Лемма 2. *Для всех $v \in V$ функции $q_v(\overline{\Lambda}_u^*, \overline{y}_u)$ монотонно возрастают по $y_{v'}$, $v' \in V_v^+$.*

Доказательство. Покажем, что для любых v и $v' \in V_v^+$ справедливо неравенство:

$$\frac{dq_v\left(\overline{\Lambda}_u^*, \overline{y}_u\right)}{dy_{v'}} > 0\,, \qquad (6)$$

где d — знак производной.

Фиксируем произвольные линии $v \in V$, $v' \in V_v^+$. Приведем ряд вспомогательных равенств (7)–(12). Как следует из определения схемы SMQMA, накопитель системы доступен для v-пакета тогда и только тогда, когда $k_v + k'_v + k''_v < a_v$, или одновременно выполняются условия

$$a_v \le k_v + k'_v + k''_v \le a_v + r_v - 1$$

и

$$\sum_{\substack{v'\in\Omega_u^+,\\ v'\ne v}} \left(k_{v'} + k'_{v'} + k''_{v'} - a_{v'}\right)^+ \le$$

$$\le R_u - 1 - \left(k_v + k'_v + k''_v - a_v\right)^+ .$$

Тогда, использовав обозначения из (1), получим:

$$q_v\left(\overline{\Lambda}_u^*, \overline{y}_u\right) = \left(\sum_{m_v=0}^{a_v-1} \sum_{\overline{k}_u\in A^{0,m_v}_{R_u,\overline{r}_u,\overline{a}_u}} p\left(\overline{c}_u, \overline{k}_u, \overline{\rho}_u^*\right) + \right.$$

$$\left. + \sum_{m_v=a_v}^{r_v+a_v-1} \sum_{\overline{k}_u\in A^{0,m_v}_{R_u-1,\overline{r}_u,\overline{a}_u}} p\left(\overline{c}_u, \overline{k}_u, \overline{\rho}_u^*\right)\right) \Big/ g(\overline{a}_u, \overline{\Lambda}_u^*, \overline{y}_u) =$$

$$= \frac{g(\overline{a}_u - \overline{1}_v, \overline{\Lambda}_u^*, \overline{y}_u)}{g\left(\overline{a}_u, \overline{\Lambda}_u^*, \overline{y}_u\right)}\,. \qquad (7)$$

Для производной функции $q_v(\overline{\Lambda}_u^*, \overline{y}_u)$ по $y_{v'}$, $v' \in V_u^+$, использовав (7), получим:

$$\frac{\partial q_v\left(\overline{\Lambda}_u^*, \overline{y}_u\right)}{\partial y_{v'}} = \frac{1}{g^2(\overline{a}_u, \overline{\Lambda}_u^*, \overline{y}_u)} \times$$

$$\times \left[g\left(\overline{a}_u, \overline{\Lambda}_u^*, \overline{y}_u\right) \frac{\partial g\left(\overline{a}_u - \overline{1}_v, \overline{\Lambda}_u, \overline{y}_u\right)}{\partial y_{v'}} - \right.$$

$$\left. - g\left(\overline{a}_u - \overline{1}_v, \overline{\Lambda}_u^*, \overline{y}_u\right) \frac{\partial g\left(\overline{a}_u, \overline{\Lambda}_u^*, \overline{y}_u\right)}{\partial y_{v'}}\right]\,. \qquad (8)$$

Из определения $g(\overline{a}_u, \overline{\Lambda}_u^*, \overline{y}_u)$, данного в (1), получим:

$$\frac{\partial g\left(\overline{a}_u, \overline{\Lambda}_u^*, \overline{y}_u\right)}{\partial y_{v'}} =$$

$$= \frac{\partial}{\partial y_{v'}} \sum_{\overline{k}\in A_{R_u,\overline{r}_u,\overline{a}_u}} \prod_{v'''\in\Omega_u^+} z_{v'''}\left(k_{v'''}, s_{v'''}, \overline{\Lambda}^*_{v'''}, \overline{y}_{v'''+}\right) =$$

$$= \sum_{v''\in\Omega_u^+} \sum_{\overline{k}\in A_{R_u,\overline{r}_u,\overline{a}_u}} \prod_{\substack{v'''\in\Omega_u^+,\\ v'''\ne v''}} z_{v'''}\left(k_{v'''}, s_{v'''}, \overline{\Lambda}^*_{v'''}, \overline{y}_{v'''+}\right) \times$$

$$\times \frac{\partial z_{v''}\left(k_{v''}, s_{v''}, \overline{\Lambda}_{v''}, \overline{y}_{v''+}\right)}{\partial y_{v'}}\,. \qquad (9)$$

Для производной функции $z_{v''}\left(k_{v''}, s_{v''}, \overline{\Lambda}^*_{v''}, \overline{y}_{v''+}\right)$ из (1) следует:

$$\left.\begin{aligned}
&\frac{\partial z_{v''}\left(k_{v''}, s_{v''}, \overline{\Lambda}^*_{v''}, \overline{y}_{v''+}\right)}{\partial \rho^*_{v''}} = \\
&\quad = \frac{k_{v''}}{\rho^*_{v''}}\, z_{v''}\left(k_{v''}, s_{v''}, \overline{\Lambda}^*_{v''}, \overline{y}_{v''+}\right); \\
&\frac{\partial z_{v''}\left(k_{v''}, s_{v''}, \overline{\Lambda}^*_{v''}, \overline{y}_{v''+}\right)}{\partial \rho'^*_{v''}} = \\
&\quad = \frac{\partial z_{v''}\left(k_{v''}, s_{v''}, \overline{\Lambda}^*_{v''}, \overline{y}_{v''+}\right)}{\partial \rho''^*_{v''}} = \\
&\quad = \frac{s_{v''}}{\rho'^*_{v''} + \rho''^*_{v''}}\, z_{v''}\left(k_{v''}, s_{v''}, \overline{\Lambda}^*_{v''}, \overline{y}_{v''+}\right).
\end{aligned}\right\} \quad (10)$$

Для производной функции $z_{v''}\left(k_{v''}, s_{v''}, \overline{\Lambda}^*_{v''}, \overline{y}_{v''+}\right)$ по $y_{v'}$ при $v' \neq v''$, $v' \notin \Omega^+_{v''+}$ получим:

$$\begin{aligned}
&\frac{\partial z_{v''}\left(k_{v''}, s_{v''}, \overline{\Lambda}^*_{v''}, \overline{y}_{v''+}\right)}{\partial y_{v'}} = \\
&\quad = \frac{\partial z_{v''}\left(k_{v''}, s_{v''}, \overline{\Lambda}^*_{v''}, \overline{y}_{v''+}\right)}{\partial \rho^*_{v''}}\, \frac{\partial \rho^*_{v''}}{\partial y_{v'}} + \\
&\quad + \frac{\partial z_{v''}\left(k_{v''}, s_{v''}, \overline{\Lambda}^*_{v''}, \overline{y}_{v''+}\right)}{\partial \rho'^*_{v''}}\, \frac{\partial \rho'^*_{v''}}{\partial y_{v'}} + \\
&\quad + \frac{\partial z_{v''}\left(k_{v''}, s_{v''}, \overline{\Lambda}^*_{v''}, \overline{y}_{v''+}\right)}{\partial \rho''^*_{v''}}\, \frac{\partial \rho''^*_{v''}}{\partial y_{v'}}. \quad (11)
\end{aligned}$$

Из определений переменных ρ^*_v, ρ'^*_v и ρ''^*_v в (1) и из (3) следуют формулы:

$$\left.\begin{aligned}
\rho^*_{v''} &= \frac{1}{y_{v''}} \sum_{l \in L_{v''}} \lambda(l) \frac{1}{1-\beta_{v''}(l)} \times \\
&\quad \times \prod_{\substack{v' \in l^+_{v''}, \\ v' \neq v''}} \left(\frac{1-\alpha_{v'}}{1-\beta_{v'}(l)} + \alpha_{v'}\right); \\
\rho'^*_{v''} &= \frac{\alpha_{v''}\tau_{v''}}{y_{v''}} \sum_{l \in L_{v''}} \lambda(l) \left(\frac{1}{1-\beta_{v''}(l)} - 1\right) \times \\
&\quad \times \prod_{\substack{v' \in l^+_{v''}, \\ v' \neq v''}} \left(\frac{1-\alpha_{v'}}{1-\beta_{v'}(l)} + \alpha_{v'}\right); \\
\rho''^*_{v''} &= \frac{\alpha_{v''}\overline{t}_{v''}}{y_{v''}} \sum_{l \in L_{v''}} \lambda(l) \times \\
&\quad \times \prod_{\substack{v' \in l^+_{v''}, \\ v' \neq v''}} \left(\frac{1-\alpha_{v'}}{1-\beta_{v'}(l)} + \alpha_{v'}\right).
\end{aligned}\right\} \quad (12)$$

Фиксируем $v'' \in \Omega^+_u$. Рассмотрим случай $v' \neq v''$, $v' \notin \Omega^+_{v''+}$. Обозначим для $l \in L_v$ через $j(v,l)$ номер линии v на пути l. Взяв производную переменных ρ^*_v, ρ'^*_v и ρ''^*_v по $y_{v'}$, из (1) и (3) получим:

$$\begin{aligned}
\frac{\partial \rho^*_{v''}}{\partial y_{v'}} &= \sum_{\substack{l:\, l \in L_{v''}, \\ v' \in l}} \frac{\partial \Lambda^*_{v''}(l)}{\mu_{v''}\left(1-\alpha_{v''}\beta_{v''}(l)\right)\partial y_{v'}} = \\
&= -\frac{1}{\mu_{v''} y_{v'}} \sum_{\substack{l:\, l \in L_{v''}, \\ v' \in l}} \frac{1}{1-\alpha_{v''}\beta_{v''}(l)} \times \\
&\quad \times \frac{1-\alpha_{v_{j(v',l)-1}}}{1-\alpha_{v_{j(v',l)-1}}\beta_{v_{j(v',l)-1}}} \frac{\lambda(l)}{y_{v_{j(v',l)}}} \times \\
&\quad \times \prod_{r=i(v'',l)}^{S_l} \left(\frac{1-\alpha_{v_r}(l)}{1-\beta_{v_r}(l)} + \alpha_{v_r}(l)\right) = \\
&= -\frac{1}{\mu_{v''} y_{v'}} \sum_{\substack{l:\, l \in L_{v''}, \\ v' \in l}} \frac{1}{1-\alpha_{v''}\beta_{v''}(l)} \times \\
&\quad \times \frac{1-\alpha_{v_{j(v',l)-1}}}{1-\alpha_{v_{j(v',l)-1}}\beta_{v_{j(v',l)-1}}} \Lambda^*_{v''}(l);
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \rho'^*_{v''}}{\partial y_{v'}} &= -\frac{\alpha_{v''}\tau_{v''}}{y_{v'}} \sum_{\substack{l:\, l \in L_{v''}, \\ v' \in l}} \frac{\beta_{v''}(l)}{1-\alpha_{v''}\beta_{v''}(l)} \times \\
&\quad \times \frac{1-\alpha_{v_{j(v',l)-1}}}{1-\alpha_{v_{j(v',l)-1}}\beta_{v_{j(v',l)-1}}} \Lambda^*_{v''}(l);
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \rho''^*}{\partial y_{v''}} &= -\frac{\alpha_{v''}\overline{t}_{v''}}{y_{v'}} \sum_{\substack{l:\, l \in L_{v''}, \\ v' \in l}} \frac{1-\beta_{v''}(l)}{1-\alpha_{v''}\beta_{v''}(l)} \times \\
&\quad \times \frac{1-\alpha_{v_{j(v',l)-1}}}{1-\alpha_{v_{j(v',l)-1}}\beta_{v_{j(v',l)-1}}} \Lambda^*_{v''}(l).
\end{aligned}$$

Подставив правые части последних трех равенств в (11), в случае $v' \neq v''$, $v' \notin \Omega^+_{v''+}$ получим:

$$\begin{aligned}
&\frac{\partial z_{v''}\left(k_{v''}, s_{v''}, \overline{\Lambda}^*_{v''}, \overline{y}_{v''+}\right)}{\partial y_{v'}} = \\
&\quad = -\frac{k_{v''}}{y_{v'}}\, z_{v''}\left(k_{v''}, s_{v''}, \overline{\Lambda}^*_{v''}, \overline{y}_{v''+}\right) \frac{1}{\mu_{v''}\rho^*_{v''}} \times \\
&\times \sum_{\substack{l:\, l \in L_{v''}, \\ v' \in l}} \frac{1}{1-\alpha_{v''}\beta_{v''}(l)} \frac{1-\alpha_{v_{j(v',l)-1}}}{1-\alpha_{v_{j(v',l)-1}}\beta_{v_{j(v',l)-1}}} \Lambda^*_{v''}(l) - \\
&\quad - \frac{s_{v''}}{y_{v'}}\, z_{v''}\left(k_{v''}, s_{v''}, \overline{\Lambda}^*_{v''}, \overline{y}_{v''+}\right) \times \\
&\times \frac{\alpha_{v''}}{\rho'^*_{v''} + \rho''^*_{v''}} \sum_{\substack{l:\, l \in L_{v''}, \\ v' \in l}} \frac{\tau_{v''}\beta_{v''}(l) + \overline{t}_{v''}\left(1-\beta_{v''}(l)\right)}{1-\alpha_{v''}\beta_{v''}(l)} \times \\
&\quad \times \frac{1-\alpha_{v_{j(v',l)-1}}}{1-\alpha_{v_{j(v',l)-1}}\beta_{v_{j(v',l)-1}}} \Lambda^*_{v''}(l). \quad (13)
\end{aligned}$$

Для случая $v' = v''$ из (12) получим:

$$\frac{\partial \rho^*_{v''}}{\partial y_{v''}} = -\frac{\rho^*_{v''}}{y_{v''}}; \qquad \frac{\partial \rho'^*_{v''}}{\partial y_{v''}} = -\frac{\rho'^*_{v''}}{y_{v''}}; \qquad \frac{\partial \rho''^*_{v''}}{\partial y_{v''}} = -\frac{\rho''^*_{v''}}{y_{v''}}.$$

Подставив правые части последних равенств и равенств (10) в (11), для случая $v' = v''$ получим:

$$\frac{dz_{v''}\left(k_{v''}, s_{v''}, \overline{\Lambda}^*_{v''}, \overline{y}_{v''+}\right)}{\partial y_{v''}} = -\frac{\left(k_{v''} + s_{v''}\right) z_{v''}\left(k_{v''}, s_{v''}, \overline{\Lambda}^*_{v''}, \overline{y}_{v''+}\right)}{y_{v''}}. \tag{14}$$

Пусть $v' \in \Omega^+_{v''*}$. Заменив переменные $\rho^*_{v''}$, $\rho'^*_{v''}$ и $\rho''^*_{v''}$ соответственно правыми частями равенств (12), для производной по $v' \in \Omega^+_{v''+}$ функции $z_{v''}\left(k_{v''}, s_{v''}, \Lambda^*_{v''}, \overline{y}_{v''+}\right)$ для случая $v' \in \Omega^+_{v''+}$ получим:

$$\frac{\partial z_{v''}\left(k_{v''}, s_{v''}, \Lambda^*_{v''}, \overline{y}_{v''+}\right)}{\partial y_{v'}} = z_{v''}\left(k_{v''}, s_{v''}, \Lambda^*_{v''}, \overline{y}_{v''+}\right)\left[\frac{s_{v''}}{\rho'^*_{v''} + \rho''^*_{v''}}\frac{\partial \rho'^*_{v''}}{\partial y_{v'}} + \frac{k_{v''}}{\rho^*_{v''}}\frac{\partial \rho^*_{v''}}{\partial y_{v'}}\right] =$$

$$= -z_{v''}\left(k_{v''}, s_{v''}, \Lambda^*_{v''}, \overline{y}_{v''+}\right)\left[\frac{s_{v''}}{\rho'^*_{v''} + \rho''^*_{v''}}\frac{\alpha_{v''}\tau_{v''}}{y_{v'}\{1 - \alpha_{v''}[1 - (1 - \delta_{v''})\, y_{v'}]\}}\sum_{\substack{l \in L_{v''},\\ v' \in l}} \Lambda^*_{v''(l)}(l) - \right.$$

$$\left. - \frac{k_{v''}}{\rho^*_{v''}}\frac{1}{\mu_{v''} y_{v'}\{1 - \alpha_{v''}[1 - (1 - \delta_{v''})\, y_{v'}]\}}\sum_{\substack{l \in L_{v''},\\ v' \in l}} \Lambda^*_{v''(l)}(l)\right] =$$

$$= -z_{v''}\left(k_{v''}, s_{v''}, \Lambda^*_{v''}, \overline{y}_{v''+}\right)\frac{1}{y_{v'}\{1 - \alpha_{v''}[1 - (1 - \delta_{v''})\, y_{v'}]\}}\sum_{\substack{l \in L_{v''},\\ v' \in l}} \Lambda^*_{v''(l)}(l)\left[\frac{k_{v''}}{\mu_{v''}\rho^*_{v''}} + \frac{\alpha_{v''}\tau_{v''}s_{v''}}{\rho'^*_{v''} + \rho''^*_{v''}}\right]. \tag{15}$$

Подставив правые части равенств (13)–(15) в (9) и использовав обозначения (5), для произвольных $v' \in V^+_v$ получим:

$$\frac{\partial g\left(\overline{a}_u, \overline{\Lambda}^*_u, \overline{y}_u\right)}{\partial y_{v'}} = -g\left(\overline{a}_u, \overline{\Lambda}^*_u, \overline{y}_u\right)\sum_{v'' \in \Omega^+_u} W_{\overline{a}_u, v'', v'}, \tag{16}$$

где

$$W_{\overline{a}_u, v'', v'} = \begin{cases} \dfrac{1}{y_{v'}}\left[\Psi_{1,v'',v'} d_{1,v'',\overline{a}_u}\left(\overline{\Lambda}^*_u, \overline{y}_u\right) + \Psi_{2,v'',v'} d_{2,v'',\overline{a}_u}\left(\overline{\Lambda}^*_u, \overline{y}_u\right)\right], & v' \neq v'', v' \notin \Omega^+_{v''+}; \\ \dfrac{1}{y_{v'''}} d_{v'',\overline{a}_u}\left(\overline{\Lambda}^*_u, \overline{y}_u\right), & v' = v''; \\ \dfrac{1}{y_{v'}}\left[\dfrac{d_{1,v'',\overline{a}_u}\left(\overline{\Lambda}^*_u, \overline{y}_u\right)}{\mu_{v''}\rho^*_{v''}} + \dfrac{\alpha_{v''}\tau_{v''} d_{2,v'',\overline{a}_u}\left(\overline{\Lambda}^*_u, \overline{y}_u\right)}{\rho'^*_{v''} + \rho''^*_{v''}}\right]\dfrac{\sum\limits_{l \in L_{v''}, v' \in l} \Lambda^*_{v''(l)}(l)}{1 - \alpha_{v''}[1 - (1 - \delta_{v''})\, y_{v'}]}, & v' \in \Omega^+_{v''+}; \end{cases}$$

$$\Psi_{1,v'',v'} = \frac{1}{\mu_{v''}\rho^*_{v''}}\sum_{\substack{l:\ l \in L_{v''},\\ v' \in l}} \frac{1}{1 - \alpha_{v''}\beta_{v''}(l)}\frac{1 - \alpha_{v_{j(v',l)-1}}}{1 - \alpha_{v_{j(v',l)-1}}\beta_{v_{j(v',l)-1}}}\Lambda^*_{v''}(l);$$

$$\Psi_{2,v'',v'} = \frac{\alpha_{v''}}{\rho'^*_{v''} + \rho''^*_{v''}}\sum_{\substack{l:\ l \in L_{v''},\\ v' \in l}} \frac{\tau_{v''}\beta_{v''}(l) + \overline{t}_{v''}\left(1 - \beta_{v''}(l)\right)}{1 - \alpha_{v''}\beta_{v''}(l)}\frac{1 - \alpha_{v_{j(v',l)-1}}}{1 - \alpha_{v_{j(v',l)-1}}\beta_{v_{j(v',l)-1}}}\Lambda^*_{v''}(l).$$

Подставив (16) в (8), получим:

$$\frac{\partial q_v\left(\overline{\Lambda}^*_u, \overline{y}_u\right)}{\partial y_{v'}} = q_v\left(\overline{\Lambda}^*_u, \overline{y}_u\right)\sum_{v'' \in \Omega^+_u} D_{\overline{a}_u, v'', v'}, \tag{17}$$

где

$$D_{\overline{a}_u,v'',v'} = \begin{cases} \dfrac{1}{y_{v'}}\left[\Psi_{1,v'',v'}\Delta_{1,v'',\overline{a}_u}\left(\overline{\Lambda}^*_u,\overline{y}_u\right) + \right. \\ \qquad \left. + \Psi_{2,v'',v'}\Delta_{2,v'',\overline{a}_u}\left(\overline{\Lambda}^*_u,\overline{y}_u\right)\right], \\ \qquad\qquad v' \neq v'', \;\; v' \neq \Omega^+_{v''+}; \\ \dfrac{1}{y_{v'}}\Delta_{v'',\overline{a}_u}\left(\overline{\Lambda}^*_u,\overline{y}_u\right), \qquad\qquad v' = v''; \\ \dfrac{1}{y_{v'}}\left[\dfrac{\Delta_{1,v'',\overline{a}_u}\left(\overline{\Lambda}^*_u,\overline{y}_u\right)}{\mu_{v''}\rho^*_{v''}} + \right. \\ \qquad \left. + \dfrac{\alpha_{v''}\tau_{v''}\Delta_{2,v'',\overline{a}_u}\left(\overline{\Lambda}^*_u,\overline{y}_u\right)}{\rho'^*_{v''}+\rho''^*_{v''}}\right]\times \\ \qquad \times \dfrac{\sum\limits_{l\in L_{v''},v'\in l}\Lambda^*_{v''(l)}(l)}{1-\alpha_{v''}\left[1-\left(1-\delta_{v''}\right)y_{v'}\right]}, \; v' \in \Omega^+_{v''+}; \end{cases}$$

$$\Delta_{v'',\overline{a}_u}\left(\overline{\Lambda}^*_u,\overline{y}_u\right) = d_{v'',\overline{a}_u}\left(\overline{\Lambda}^*_u,\overline{y}_u\right) - d_{v'',\overline{a}_u-1_v}\left(\overline{\Lambda}^*_u,\overline{y}_u\right);$$

$$\Delta_{1,v'',\overline{a}_u}\left(\overline{\Lambda}^*_u,\overline{y}_u\right) = d_{1,v'',\overline{a}_u}\left(\overline{\Lambda}^*_u,\overline{y}_u\right) - d_{1,v'',\overline{a}_u-1_v}\left(\overline{\Lambda}^*_u,\overline{y}_u\right);$$

$$\Delta_{2,v'',\overline{a}_u}\left(\overline{\Lambda}^*_u,\overline{y}_u\right) = d_{2,v'',\overline{a}_u}\left(\overline{\Lambda}^*_u,\overline{y}_u\right) - d_{2,v'',\overline{a}_u-1_v}\left(\overline{\Lambda}^*_u,\overline{y}_u\right).$$

Отметим, что согласно лемме 1 $\Delta_{v'',\overline{a}_u}\left(\overline{\Lambda}^*_u,\overline{y}_u\right) > 0$, $\Delta_{1,v'',\overline{a}_u}\left(\overline{\Lambda}^*_u,\overline{y}_u\right) > 0$ и $\Delta_{2,v'',\overline{a}_u}\left(\overline{\Lambda}^*_u,\overline{y}_u\right) > 0$. Тогда, так как $0 \le \alpha_{v''} \le 1$, $0 \le \delta_{v''} < 1$, $0 < y_{v''} \le 1$, $\Psi_{1,v'',v'} > 0$, $\Psi_{2,v'',v'} > 0$ для всех $v', v'' \in V$ таких, что $v', v'' \in l$ хотя бы для одного $l \in L$, из (17) получим неравенство (6) и, следовательно, доказательство леммы 2.

Продолжим доказательство теоремы. Из определения последовательности $\overline{y}[n]$, $n \ge 0$, и из леммы 2 следует, что $\overline{y}[n+1] < \overline{y}[n]$, $n \ge 0$. Существование предела последовательности $\overline{y}[n]$, $n \ge 0$, следует непосредственно из свойства монотонности и ограниченности этой последовательности.

Пусть $\overline{y}' = \{y'_v \in (0,\,1],\; v \in V\}$ — положительное решение системы уравнений (4), $\overline{\Lambda}'_u$ — значение переменной $\overline{\Lambda}^*_u$ при $\overline{y}'$, $\overline{\Lambda}^*_u[n]$ — значение переменной $\overline{\Lambda}^*_u$ при $\overline{y}_u[n]$. Очевидно, что $y'_v < 1$, $v \in V$, так как $q_v\left(\overline{\Lambda}^*_u,\overline{y}_u\right) < 1$ при любых положительных $\overline{y}$. Пусть $\overline{y}[n] > \overline{y}'$ для некоторого $n \ge 0$ (существование такого n вытекает из того, что $y_v[0] = 1$ и $y'_v < 1$, $v \in V$). Тогда, как следует из леммы 2, $y_v[n+1] = q_v\left(\overline{\Lambda}^*_u[n],\overline{y}_u[n]\right) > q_v\left(\overline{\Lambda}'_u,\overline{y}'_u\right) = y'_v$ для каждой линии $v \in V$, т. е. последовательность $\overline{y}[n]$, $n \ge 0$, ограничена снизу величиной $\overline{y}'$. Значит, существуют пределы $\lim\limits_{n\to\infty} y_v[n] = y^0_v \ge y'_v > 0$ для всех $v \in V$. Так как $\overline{\Lambda}^*_u$, $q_v\left(\overline{\Lambda}^*_u,\overline{y}_u\right)$, $v \in V$, — непрерывные по $y_{v'}$, $v' \in V^+_v$, функции, то можно написать:

$$\lim_{n\to\infty} q_v\left(\overline{\Lambda}^*_u[n],\overline{y}_u[n]\right) = q_v\left(\overline{\Lambda}^0_u,\overline{y}^0_u\right) = y^0_v,$$

где $\overline{\Lambda}^0_u$ — значение переменной $\overline{\Lambda}^*_u$ при $y^0_{v'}$, $v' \in V^+_v$, $u = v^-$, т. е. $\overline{y}^0 = \{y^0_v \in (0,1),\; v \in V\}$ — положительное решение системы уравнений (4).

Пусть теперь $\lim\limits_{n\to\infty} y_v[n] = y^0_v > 0$ для всех $v \in V$. Тогда, так как $\overline{\Lambda}^*_u$, $q_v\left(\overline{\Lambda}^*_u,\overline{y}_u\right)$, $v \in V$, — непрерывные по $y_{v'}$, $v' \in V^+_v$, функции, выполняется $q_v\left(\overline{\Lambda}^0_u,\overline{y}^0_u\right) = y^0_v$, т. е. $\overline{y}^0 = \{y^0_v \in (0,\,1),\; v \in V\}$ — положительное решение системы уравнений (4).

Следствие 1. Система (4) не имеет положительного решения тогда и только тогда, когда $\lim\limits_{n\to\infty} y_v[n] = y^*_v = 0$ хотя бы для одного $v \in V$.

Следствие 2. Первичные потоки — реализуемые (т. е. интенсивности первичных входных потоков равны интенсивностям соответствующих выходных потоков) тогда и только тогда, когда $\lim\limits_{n\to\infty} y_v[n] = y^*_v > 0$ для всех $v \in V$.

4 Алгоритм и пример расчета

Предлагается следующий алгоритм вычисления загруженности линий $\overline{\rho}^*_v$, $v \in V$, и вероятностей блокировки пакетов π_v, $v \in V$, основывающийся на изложенном выше методе поиска решения системы (4) (методе простой итерации). Для обозначения значений, вычисляемых на k-м шаге алгоритма, к обозначениям соответствующих параметров приписывается знак $[k]$.

Шаг 1. *Инициализация.* Используя рекуррентную формулу (см. (3))

$$\Lambda^*_v[0](l) = \lambda(l)\prod_{v'\in l^+_v}\frac{1-\alpha_{v'}\delta_{v'}}{1-\delta_{v'}},\; l \in L_v,\; v \in V,$$

и соответствующие формулы из (1), вычислить начальные значения параметров $\overline{\rho}^*_v$, y_v, $v \in V$: $\overline{\rho}^*_v[0] = (\rho^*_v[0], \rho'^*_v[0], \rho''^*_v[0])$, $y_v[0] = 1$.

Положить $k = 1$.

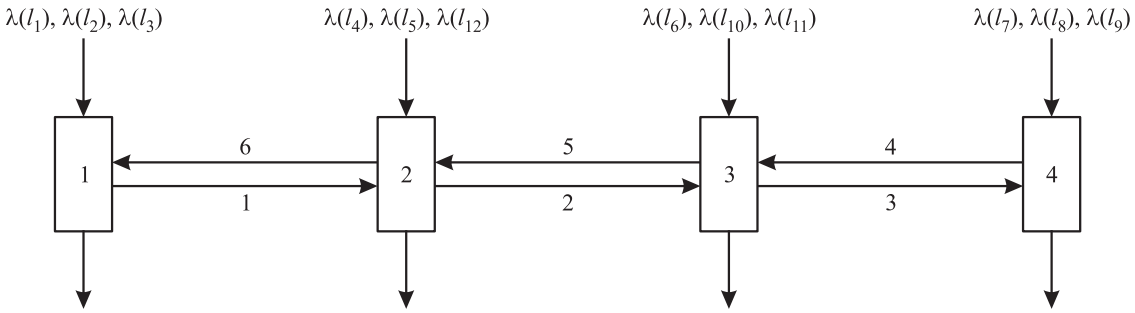


Рис. 2 Структура сети

Шаг k **(**$k \geq 1$**).**

1. *Вычисление вероятностей блокировки*. Используя значения параметров $\overline{\rho}_v^*[k-1]$, $v \in V$, с помощью соответствующих формул из (1) вычислить $y_v[k] = 1 - \pi_v[k]$, $v \in V$. При этом используется метод свертки Базена, позволяющий производить рекуррентные вычисления [3].

2. *Проверка условий останова алгоритма*. Если хотя бы для одной $v \in V$ выполняется условие $y_v[k] < \varepsilon$, где $\varepsilon > 0$ — значение требуемой точности результатов, алгоритм завершает работу, выдав сообщение: «Первичные потоки заявок не реализуемы».

3. *Вычисление значения параметра* $\overline{\rho}_v^*[k]$ с помощью рекуррентной формулы (3) и формул (1), подставив $y_v = y_v[k]$, $v \in V$.

4. *Проверка условий останова алгоритма*. Если хотя бы для одной $v \in V$ выполняется условие

$$\frac{|\rho_v^*[k] - \rho_v^*[k-1]|}{\rho_v^*[k]} > \varepsilon\,,$$

то перейти к шагу $k+1$, положив k равным $k+1$, иначе алгоритм завершает работу.

По завершении алгоритма либо выявляется, что система уравнений не имеет положительного решения (первичные потоки не реализуемы), либо вычисляются загруженности линий $\overline{\rho}_v^* = (\rho_v^*, \rho_v'^*, \rho_v''^*)$ и вероятности блокировки узлов для v-пакетов π_v, $v \in V$. Далее с помощью известных формул (в том числе формулы (1)) и полученных значений параметров $\lambda_v^*(l)$, $\overline{\rho}_v^*$, π_v, $l \in L$, $v \in V$, могут быть вычислены другие характеристики сети: среднее значение задержки пакета в узлах, среднее число повторов пакета в узлах и из источника, среднее число пакетов, находящихся в сети и в ожидании повтора из источника, и др.

Число арифметических операций, выполняемых на одном шаге алгоритма при использовании метода свертки Базена, имеет порядок

$$Q = K \sum_{u \in U} \left|\Omega_u^+\right| R_u + \sum_{l \in L} |l|\,,$$

где K — константа; $|l|$ — длина пути l.

В качестве примера использования разработанного алгоритма проведены расчеты характеристик сети с четырьмя узлами, топология которой задается графом, показанным на рис. 2. Фиксированы следующие входные данные:

$$U = \{1,2,3,4\}\,;\; V = \{1,2,3,4,5,6\}\,;$$
$$L = \{l_1, \ldots, l_{12}\}\,,$$
$$l_1 = \{1\},\; l_2 = \{1,2\}\,,\; l_3 = \{1,2,3\}\,,$$
$$l_4 = \{2\}\,,\; l_5 = \{2,3\}\,, l_6 = \{3\}\,,\; l_7 = \{4\}\,,$$
$$l_8 = \{4,5\}\,,\; l_9 = \{4,5,6\}\,,\; l_{10} = \{5\}\,,\; l_{11} = \{5,6\}\,,$$
$$l_{12} = \{6\}\,;\; R_1 = R_4 = 20\,;\; R_2 = R_3 = 25\,;$$
$$r_1 = r_4 = 20\,;\; r_2 = r_3 = r_5 = r_6 = 25\,;$$
$$a_i = 0\,;\; \mu_i = 1\,;\; \delta_i = 0{,}001\,;\; c_i = c_j\,;$$
$$\alpha_i = \alpha_j\,,\;\; i,j = 1,\ldots,10\,;$$
$$\lambda(l_i) = \lambda(l_j)\,, i,j = 1,\ldots,12\,.$$

При расчете $\overline{N}_{\text{повт}}$ — среднего суммарного числа пакетов, ожидающих повтора, использована формула:

$$\overline{N}_{\text{повт}} = \sum_{l \in L} \left(\Lambda_{v_{1(l)}}^*(l) - \lambda(l)\right) t_{\text{повт}}(l)\,,$$

где $v_{1(l)}$ — первая линия в составе пути l, $t_{\text{повт}}(l)$ — интервал времени ожидания пакетом потока $\lambda(l)$ повторного поступления с момента отказа в передаче, $t_{\text{повт}}(l) = 10$.

В таблице и на рис. 3 приведены зависимости от интенсивности первичных потоков $\lambda(l)$, $l \in L$, среднего числа пакетов в сети (ряды с нечетными номерами), суммарного среднего числа пакетов в сети и в источниках в ожидании повтора (ряды с четными номерами).

Зависимости среднего числа пакетов от интенсивности первичных потоков

Номер ряда	a_i	c_i	$\lambda(l)$									
			0,05	0,25	0,45	0,65	0,85	1,05	1,25	1,45	1,6	1,8
1	1	8	4,805	24,024	43,277	65,483	∞	∞	∞	∞	∞	∞
2	1	8	4,805	24,024	43,327	70,086	∞	∞	∞	∞	∞	∞
3	0	8	1,602	8,008	14,414	20,828	27,283	33,889	40,917	49,126	59,651	∞
4	0	8	1,612	8,058	14,504	20,958	27,455	34,15	41,813	54,234	88,368	∞
5	0	8	1,603	7,999	14,411	20,814	27,325	33,995	40,964	49,215	∞	∞
6	0	8	1,619	8,083	14,566	21,044	27,632	34,465	42,6	55,123	∞	∞
7	1	1	3,408	17,13	31,036	45,252	61,404	∞	∞	∞	∞	∞
8	1	1	3,408	17,130	31,037	45,390	64,929	∞	∞	∞	∞	∞
9	0	1	0,204	1,114	2,207	3,546	5,229	7,42	10,415	14,83	22,201	34,70561
10	0	1	0,214	1,164	2,297	3,676	5,4	7,63	10,669	15,214	23,873	46,70904

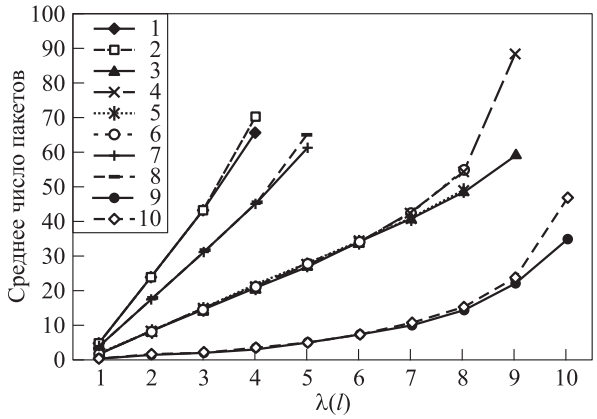


Рис. 3 Зависимости среднего числа пакетов от интенсивности первичных потоков: 1–10 — ряды 1–10

5 Заключение

Заметим, что погрешность, которую вносит алгоритм, ограничивается заранее задаваемой малой величиной $\varepsilon > 0$ и достоверность результатов работы алгоритма зависит только от адекватности модели сети. Результаты исследований разветвленных сетей показывают, что использованные выше в модели упрощающие предположения о пуассоновских входных потоках и независимости времен обслуживания в узлах на практике подтверждаются (см., например, [3, 4, 7–10]). Отметим также, что анализ результатов вычислительных экспериментов, проведенных с использованием имитационных моделей, показал слабую зависимость вероятностей блокировки узлов, среднего числа пакетов в сети и среднего числа пакетов, ожидающих повтора в источниках, от вида функции распределения интервала ожидания повторной передачи пакета из источника.

Открытым остается вопрос единственности решения системы уравнений (4). Для некоторых частных случаев вопрос решается положительно. Приведем доказательство единственности решения системы (4) для следующего частного случая сети. Рассмотрим сеть, в которой передача пакетов по каждой линии происходит только в одном направлении и используется полнодоступная схема управления буферами (CS). Заметим, что для схемы CS при всех $v \in \Omega_u^+$ верны $\pi_v = \pi_u$, π_u — вероятность блокировки узла u

Пусть M_0 — множество абонентских узлов-стоков (абонентов-адресатов): M_i — множество узлов связи $u' \notin \bigcup\limits_{j=1}^{i-1} M_j$ таких, что для всех $v \in \Omega_{u'}^+$ выполняется $v^+ \in \bigcup\limits_{j=1}^{i-1} M_j$, $i = 1, \ldots, l_{\max}$, где $l_{\max} = \max\limits_{l \in L} S_l$ — длина максимально протяженного пути на множестве L. Легко показать, что $M_1, \ldots, M_{l_{\max}}$ обладают свойствами:

(1) $M_i \frown M_j = \Theta$, $i, j = 1, \ldots, l_{\max}$, $i \neq j$;

(2) $\bigcup\limits_{j=1}^{l_{\max}} M_j = U$;

(3) максимальная длина пути (число линий, входящих в путь), соединяющего $u \in M_i$ с абонентом-адресатом, равна i;

(4) множество $\bigcup\limits_{j=1}^{i-1} M_j$ — множество всех различных узлов, следующих после любого $u \in M_i$ по направлению к адресату на путях, содержащих u.

Пусть $u \in M_1$. Тогда для любой $v \in \Omega_u^+$ верно $\Lambda_v^* = \sum\limits_{l \in L_v} \lambda(l)/y_u$ (так как v — абонентская линия). Так как при схеме CS для всех $v \in \Omega_u^+$ выполняется $y_v = y_u = 1 - \pi_u$, то уравнение (4) эквивалентно уравнению:

$$y_u = q_u\left(\overline{\Lambda}_u^*, y_u\right) \tag{18}$$

и имеет одну неизвестную переменную y_u.

Введем параметры:

$$\Lambda_v(l) = \lambda(l) \prod_{v' \in l_v} \left(\frac{1-\alpha_{v'}}{1-\beta_{v'}} + \alpha_{v'} \right) ;$$

$$\Lambda_v = \sum_{l \in L_v} (l) ; \quad \delta'_v = \alpha_v \beta_v ;$$

$$t'_v = \frac{\overline{t}_v \alpha_v (1-\beta_v)}{1-\alpha_v \beta_v} , \quad v \in \Omega_u^+ .$$

Как видим, введенные параметры для любого $u \in \in M_i$ однозначно определяются переменными $y_{u'}$, $u' \in \bigcup\limits_{j=1}^{i-1} M_j$ (следует из приведенных выше выражений для введенных параметров и из свойства 4 множеств M_j).

Заметим, что (18) сводится к уравнению (3.3) из работы [10], если в нем положить $\alpha_v = 1$, $\lambda_v = \Lambda_v$, $\delta_v = \delta'_v$, $t_v = t'_v$, $v \in \Omega_u^+$. Тогда из теоремы 1 работы [10] следует, что в рассматриваемом случае уравнение (18) имеет единственное решение.

Фиксируем некоторое $i \in \overline{2, \ldots, l_{\max}}$. Предположим, что для всех $u' \in \bigcup\limits_{j=1}^{i-1} M_j$ существуют единственные решения $y_{u'}$ уравнений (4) (выражения правых частей этих уравнений не включают переменные $y_u \notin \bigcup\limits_{j=1}^{i-1} M_j$). Тогда для случая любого $u \in M_i$ аналогично случаю $u \in M_1$ получим уравнение вида (18), которое имеет единственное решение. Следовательно, для любого $u \in \bigcup\limits_{j=1}^{l_{\max}} M_j$ уравнение (18) имеет единственное решение. Так как $\bigcup\limits_{j=1}^{l_{\max}} M_j = U$ (см. выше свойство 2 множеств M_j), то получаем доказательство единственности решения системы (4) для рассматриваемого случая.

Отметим также следующие свойства разработанного алгоритма:

- при реализуемых первичных потоках последовательность $\overline{y}[n], n \geq 0$, сходится к положительному решению системы уравнений (4);
- если последовательность $\overline{y}[n]$, $n \geq 0$, сходится к неположительному вектору $\overline{y}^0$, то входные первичные потоки не реализуемы в сети;
- при реализуемых первичных потоках алгоритм вычисляет вероятности блокировки узлов и загруженности узлов и каналов связи с приемлемой для предварительного анализа сети точностью (относительная погрешность вероятности блокировки $\sim 0{,}1$) .

Алгоритм рекомендуется использовать для оценки эффективности решений по выбору канальных и вычислительных ресурсов, схем управления буферами и маршрутизации на этапе предварительного проектирования современных сетей коммутации пакетов.

Литература

1. *Kamoun F., Kleinrock L.* Analysis of shared finite storage in a computer networks node environment under general traffic conditions // IEEE Trans. on Communications, 1980. Vol. 28. No. 7. P. 992–1003.
2. *Ефимушкин В. А., Ледовских Т. В., Салькова М. В.* Механизмы управления трафиком в сетях ATM // Электросвязь, 2003. № 1. С. 39–41.
3. *Башарин Г. П., Бочаров П. П., Коган Я. А.* Анализ очередей в вычислительных сетях. — М.: Наука, 1989.
4. *Вишневский В. М.* Теоретические основы проектирования компьютерных сетей. — М.: Техносфера, 2003.
5. *Simon S. L.* Store-and-forward buffer requirements in a packet switching network // IEEE Trans. on Communications, 1976. Vol. COM-24. No. 4. P. 394–403.
6. *Kelly F. P.* Blocking probabilities in large circuit-switched networks // Advances Appl. Probability, 1986. Vol. 18. P. 473–505.
7. *Агаларов Я. М.* Приближенный метод вычисления характеристик узла телекоммуникационной сети с повторными передачами // Информатика и её применения, 2009. Т. 3. Вып. 2. С. 2–10.
8. *Степанов С. Н.* Основы телетрафика мультисервисных сетей. — М.: Эко-Трендз, 2010.
9. *Agalarov Y.* Algorithm of nodes load estimation in the network with repetitions from source and static buffer management scheme // 2010 Congress (International) on Ultra Modern Telecommunications and Control Systems (ICUMT) Proceedings. — M., 2010. P. 1073–1077.
10. *Агаларов Я. М.* Об одном численном методе вычисления стационарных характеристик узла коммутации с повторными передачами // Автоматика и телемеханика, 2011. № 1. С. 95–106.

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2012. Т. 6. Вып. 2. С. 14–21

ЗАДАЧИ АНАЛИЗА И ОПТИМИЗАЦИИ ДЛЯ МОДЕЛИ ПОЛЬЗОВАТЕЛЬСКОЙ АКТИВНОСТИ. ЧАСТЬ 3. ОПТИМИЗАЦИЯ ВНЕШНИХ РЕСУРСОВ

А. В. Босов[1]

Аннотация: Статья завершает исследование модели описания активности пользователей, предложенной автором ранее, и основанных на ней задач оптимизации распределения вычислительных ресурсов. Сформулирована и решена задача квадратичной оптимизации распределения «внешних» ресурсов, используемых информационной системой. Предложены субоптимальные алгоритмы оптимизации.

Ключевые слова: информационная система; система управления базами данных; стохастическая система наблюдения; квадратичный критерий

1 Введение

Предложенная в работе [1] математическая модель описания пользовательской активности рассматривалась в качестве источника для постановки задач оптимизации распределения вычислительных ресурсов, используемых некоторой информационной системой при обслуживании запросов пользователей.

В [2] стохастическая динамическая система наблюдения, описывающая эволюцию показателя пользовательской активности (текущее число пользователей x_t) моделью случайного процесса с переключениями, порождаемыми значениями показателя, и линейными наблюдениями за ним (число выполненных команд y_t), дополнена квадратичным функционалом качества, задающим штраф за использование программой «внутренних» вычислительных ресурсов.

Целевой функционал сформирован в результате анализа алгоритма работы программного обеспечения Информационного веб-портала [3, 4], «внутренним» ресурсом которого является пул, поддерживаемый с целью параллельного выполнения поступающих запросов.

Внутренний характер рассмотренного в [2] вида ресурса объясняется тем, что управление им полностью подконтрольно программной системе, а выбор стратегии оптимизации никак не влияет на состояние внешней среды (активность пользователей, обслуживающие системы). В результате получена оригинальная постановка задачи оптимизации: ни состояние, ни наблюдения не зависят непосредственно от выбранной стратегии, обратную связь обеспечивает только минимизируемый целевой функционал.

В данной статье использована та же модель активности пользователей и аналогичный подход к оптимизации, но уже с целью управления распределением вычислительных ресурсов «внешних», т. е. обслуживающих систем. Как отмечалось в [2], предложить осмысленные постановки для таких задач существенно труднее, чем для «внутренних» ресурсов. К последним можно отнести практически любой программный объект, создаваемый в процессе функционирования рассматриваемого программного обеспечения (а значит, и полностью им контролируемый). Для «внешних» ресурсов, которые любой программой, несомненно, используются во множестве, нужно иметь доступные средства управления их выделением извне, а возможностей такого рода известно немного.

В процессе работы Информационного веб-портала используется такой «внешний» ресурс, и поддерживается он применяемой системой управления базами данных (СУБД). Одним из реализуемых порталом сервисов является личный кабинет пользователя, поддерживаемый портальной подсистемой представления в составе средств персонализации. Эффективность функционирования этой подсистемы хотя и не столь критична, как у пула запросов, но также допускает постановку задачи оптимизации, решение которой вносит свой вклад в повышение качества работы портала в целом.

Функциональность, реализованная в связи с поддержкой личного кабинета, обеспечивает пользователя возможностью сохранения результатов выполнения пользовательской команды. Инфор-

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

мация, отобранная по заданным критериям, может быть перенесена в индивидуальное хранилище, и в дальнейшем сохраненные ранее результаты могут быть проанализированы вновь в рамках интерфейса личного кабинета.

Совокупность персональных хранилищ данных образует информационный источник (пользовательскую базу данных), поддерживаемый в составе собственного хранилища портала, в связи с чем возникают традиционные задачи его сопровождения: резервного копирования и восстановления, выделения дисковых ресурсов и сборки мусора и т. п. Предметом оптимизации здесь является процедура обслуживания дискового пространства, выделяемого под размещение пользовательской базы данных. Накладные расходы, влияющие на функционирование портала в целом, в этой связи не слишком велики, если размер пользовательской базы данных мал. Но с ростом этого размера, что рано или поздно неизбежно происходит, могут сказываться уже существенно.

Отметим следующие причины, по которым в Информационном веб-портале обслуживание пользовательской базы данных выполняется особым образом:

- анализ потребностей потенциальных пользователей привел к отказу от применения традиционных принудительных ограничений на пространство, выделяемое под персональные данные конкретного пользователя, по причинам (а) количественных характеристик пользовательской аудитории; (б) неконтролируемого разнообразия функциональных потребностей; (в) недопустимости принудительного вмешательства в ход выполняемой пользователем текущей работы из-за нехватки ресурсов;
- единственным приемлемым способом освобождения персональной базы данных от устаревших сведений признаны уведомительные сообщения, формируемые исходя из временны́х меток, устанавливаемых на сохраненные данные;
- традиционные для современных СУБД механизмы автоматического расширения дискового пространства при его нехватке на заданный процент от текущего размера неприемлемы: при малом размере базы данных операция расширения будет выполняться слишком часто, при большом — будет выделено неоправданно много места;
- аналогичная причина заставила отказаться от использования механизма сжатия, выполняемого либо автоматически, либо по команде администратора, — как будет пояснено далее, портальная функциональность предполагает наличие определенного (но не слишком большого) свободного пространства в пользовательской базе данных в любой момент времени, сжатие же предполагает использование совместно с автоматическим расширением.

Цель обслуживания пользовательской базы данных — обеспечение компактного хранения уже отобранных пользователями данных (а значит, и эффективная индексация, и доступ) и поддержка небольшого, но приемлемого свободного объема, достаточного для удовлетворения потребностей пользователей в каждый момент времени. Попутно также удается минимизировать деятельность администратора по сопровождению хранилищ личных кабинетов.

2 Используемые обозначения

Далее в работе будут использованы следующие обозначения:

- $\triangleq$ — равенство по определению;
- $\mathbf{M}[x]$ и $\mathbf{M}[x|\mathfrak{J}]$ — соответственно безусловное математическое ожидание случайной величины x и условное математическое ожидание x относительно σ-алгебры $\mathfrak{J}$;
- x^{T} — операция транспонирования вектора (матрицы) x;
- $\mathrm{col}\,(x_1,\ldots,x_n) \triangleq (x_1,\ldots,x_n)^{\mathrm{T}}$ — вектор-столбец с элементами $x_1,\ldots,x_n$;
- $\mathrm{row}\,(x_1,\ldots,x_n) \triangleq (x_1,\ldots,x_n)$ — вектор- строка с элементами $x_1,\ldots,x_n$;
- $\mathfrak{J}_t^y \triangleq \sigma\{y_\tau, \tau \le t\}$ — σ-алгебра, порожденная наблюдениями y_τ, $\tau \le t$;
- $\overline{\psi}_x(x,t,j)$, $j = 0,1,\ldots,$ — условная плотность вероятности x_{t+j} относительно σ-алгебры $\mathfrak{J}_{t-1}^y$;
- $\hat{\psi}_x(x,t)$ — условная плотность вероятности x_t относительно $\mathfrak{J}_t^y$.

3 Модель распределения «внешних» ресурсов портала

Для описания процесса изменения размера пользовательской базы данных воспользуемся той же моделью, что и в [1, 2], т. е. будем считать, что показатель пользовательской активности x_t за интервал наблюдения $(t-1;t]$ описывается следующим разностным стохастическим уравнением:

$$x_t = a\Theta(x_{t-1})x_{t-1} + q\Theta(x_{t-1}) + \\ + b\Theta(x_{t-1})v_t\,,\ \ t = 1, 2, \ldots\,, \quad (1)$$

предполагая, что область значений x_t разбита на непересекающиеся интервалы Δ_k:

$$-\infty = a_1 < a_2 < \cdots < a_n < a_{n+1} = +\infty\,;$$
$$\Delta_k = (a_k, a_{k+1}],\ k = 1, \ldots, n-1;\ \Delta_n = (a_n, +\infty)$$

и текущий режим пользовательской активности задан индикаторной функцией $\theta(x)$:

$$\left.\begin{aligned} \Theta(x) &= \operatorname{col}\left(I_{\Delta_1}(x), \ldots, I_{\Delta_n}(x)\right)\,; \\ I_{\Delta_k}(x) &= \begin{cases} 1\,, & \text{если } x \in \Delta_k\,; \\ 0\,, & \text{если } x \notin \Delta_k\,; \end{cases} \\ a &= \operatorname{row}\left(a_1, \ldots, a_n\right)\,; \\ q &= \operatorname{row}\left(q_1, \ldots, q_n\right)\,; \\ b &= \operatorname{row}\left(b_1, \ldots, b_n\right)\,. \end{aligned}\right\} \quad (2)$$

В качестве наблюдений y_t будем использовать объем пользовательских данных, помещенных в хранилище в момент времени t. Уравнение для y_t получим из следующих соображений. Заметим, что содержимое хранилища в момент t должно складываться из данных, внесенных пользователями до момента $t-1$ и не удаленных за интервал наблюдения $(t-1; t]$, и новых данных, пополнивших хранилище за последний интервал наблюдения. Будем предполагать, что часть сохраняемых от шага к шагу данных пропорциональна их размеру и составляет в среднем dy_t, т. е. $d \cdot 100\%$, $0 < d < 1$ пользовательской информации, имеющейся в хранилище в момент $t-1$, остается в нем и в момент t, а $(1-d)\cdot 100\%$ — удаляется. Также будем предполагать, что объем новых данных, пополняющих хранилище за время $(t-1; t]$, пропорционален числу активных пользователей и составляет в среднем cx_t, где параметр c определяет средний объем информации, размещаемой в хранилище персональных данных одним пользователем за интервал наблюдения $(t-1; t]$. Таким образом, объем пользовательских данных в хранилище можно описывать следующим уравнением:

$$y_t = dy_{t-1} + cx_t + \sigma w_t\,; \quad y_0 = 0\,, \quad (3)$$

где w_t — возмущение, моделирующее отклонения размера фактически сохраняемых и фактически добавляемых данных от заданных средних уровней; σ — среднее квадратическое отклонение этого возмущения. Далее предполагается, что $\{w_t\}$ — стандартный дискретный белый шум второго порядка.

Собственно же размер пользовательской базы данных z_t определяется только тем управляющим воздействием u_t, которое будет применено непосредственно к пользовательской базе данных средствами администрирования СУБД, а именно: в момент времени $t-1$ должно быть принято решение о расширении (уменьшении) выделенного под хранилище персональных данных дискового пространства на величину u_t. Таким образом, выход z_t системы наблюдения (1)–(3) вычисляется в момент $t-1$ как сумма z_{t-1} и управления u_t, которое формируется на основании всех наблюдений, выполненных к моменту $t-1$:

$$z_t = z_{t-1} + u_t\,; \quad z_0 = G_0\,. \quad (4)$$

Целью оптимизации является определение подходящего размера пользовательской базы данных, позволяющего разместить в ней всю пользовательскую информацию, накопленную к моменту $t-1$, а также оставить свободное место некоторого объема G_t, достаточного для размещения вновь поступающей за время $(t-1; t]$ пользовательской информации. Для достижения этого в целевую функцию следует включить следующие слагаемые:

$$\mathbf{M}\left[(z_t - G_t - y_t)^2 + u_t^2\right]\,, \quad (5)$$

т. е. определять стратегию оптимизации, штрафуя за разницу между предполагаемым $(z_t - G_t)$ и фактическим y_t объемом пользовательских данных в момент времени t, а также за величину управляющего воздействия u_t. Отметим, что в (5) не включено слагаемое, позволяющее учесть штраф за суммарный размер хранилища. Включение такого слагаемого z_t^2 неминуемо привело бы к ограничению на размер пользовательской базы, что в условиях портала, как отмечалось выше, неприемлемо. Однако нетрудно видеть, что полученный далее результат легко может быть распространен и на данный вид штрафа.

Для придания (5) окончательного вида предположим, что задан горизонт оптимизации N (например, сутки или неделя) и дополним выбранные слагаемые весовыми коэффициентами. Окончательно получаем целевую функцию следующего вида:

$$\mathbf{J}\left(u_0, \ldots, u_N\right) = \\ = \sum_{t=1}^{N} \mathbf{M}\left[S_t^{(1)}\left(z_t - G_t - y_t\right)^2 + S_t^{(2)} u_t^2\right]\,. \quad (6)$$

4 Формальная постановка и решение задачи управления размером пула

Всюду далее будем предполагать, что $\{v_t\}$ из (1) — стандартный дискретный белый шум

в узком смысле, сечения которого имеют плотность вероятности $\varphi_v(\cdot)$; x_0 — случайная величина, имеющая плотность вероятности $\psi_0(\cdot)$; $\{w_t\}$ из (3) — стандартный дискретный белый шум в узком смысле, сечения которого имеют плотность вероятности $\varphi_w(\cdot)$; $\{v_t\}$, x_0, $\{w_t\}$ независимы в совокупности; $\mathbf{M}\left[v_t^2\right] < \infty$; $\mathbf{M}\left[x_0^2\right] < \infty$; $\mathbf{M}\left[w_t^2\right] < \infty$; параметры b_k, $k = 1, \ldots, n$ и σ неотрицательны.

Будем предполагать также, что параметры целевого функционала (6) $S_t^{(1)}$, $S_t^{(2)}$, G_t — известные неотрицательные функции t, класс допустимых управлений U_t содержит все $\mathfrak{J}_{t-1}^y$-измеримые функции u_t : $\mathbf{M}[u_t^2] < \infty$. Отметим, что отсюда вытекает $\mathfrak{J}_{t-1}^y$-измеримость процесса z_t. Таким образом, целью оптимизации является поиск закона изменения u_t^* размера хранилища пользовательских данных, удовлетворяющего потребности в ресурсах, описываемых наблюдениями y_t, с минимизацией затрат на управляющее воздействие на текущем шаге и на всех последующих шагах вплоть до заданного горизонта N:

$$\operatorname{col}\left(u_0^*, \ldots, u_N^*\right) = \\ = \underset{(u_0,\ldots,u_N)\in U_0\times\cdots\times U_N}{\arg\min} \mathbf{J}\left(u_0, \ldots, u_N\right). \quad (7)$$

Теорема. *Пусть для целевого функционала* (6) *выполнено:* $S_t^{(1)} + S_t^{(2)} > 0$, $1 \le t \le N$. *Тогда решение* u_t^* *задачи оптимизации* (7) *существует и определяется соотношением:*

$$u_t^* = \frac{1}{R(t)}\left(\sum_{j=0}^{N-t} Q_j(t)\left(\overline{y}_{t+j,t-1} + G_{t+j}\right) - \right. \\ \left. - P(t)z_{t-1}\right). \quad (8)$$

Здесь

$$R(t) = S_t^{(2)} + P(t)\,,\ 1 \le t \le N\,; \\ P(t) = S_t^{(1)} + L(t+1)\,,\ 1 \le t \le N\,,$$

где

$$L(t) = L(t+1) + S_t^{(1)} - \frac{P^2(t)}{R(t)}\,,\ 1 \le t \le N\,, \\ L(N+1) = 0\,; \\ Q_j(t) = Q_{j-1}(t+1) - Q_{j-1}(t+1)\frac{P(t+1)}{R(t+1)}\,; \\ Q_0(t) = S_t^{(1)}\,,\ 1 \le t \le N\,,\ 0 \le j \le N-t\,;$$

$\overline{y}_{t+j,t-1}$ — оптимальные в среднем квадратическом прогнозы наблюдений y_{t+j} по наблюдениям y_τ, $\tau \le t-1$.

Доказательство. Для решения задачи оптимизации (7) воспользуемся методом динамического программирования [5, 6]. Обозначим через

$$B(t) \triangleq \\ \triangleq \min_{(u_t,\ldots,u_N)\in U_t\times\cdots\times U_N} \sum_{\tau=t}^{N} \mathbf{M}\left[S_\tau^{(1)}\left(z_\tau - G_\tau - y_\tau\right)^2 + \right. \\ \left. + S_\tau^{(2)} u_\tau^2 | \mathfrak{J}_{t-1}^y\right]$$

функцию Беллмана. При $t = N$ утверждение теоремы очевидно, так как из выражения

$$B(N) = \\ = \min_{u_N\in U_N} \mathbf{M}\left[S_N^{(1)}\left(z_N - G_N - y_N\right)^2 + S_N^{(2)} u_N^2 | \mathfrak{J}_{N-1}^y\right]$$

после очевидных преобразований с учетом $\mathfrak{J}_{N-1}^y$-измеримости функции u_N, равенства $z_t = z_{t-1} + u_t$, а также обозначения $\overline{y}_{N,N-1} = \mathbf{M}\left[y_N | \mathfrak{J}_{N-1}^y\right]$ получается

$$u_N^* = \arg\min_{u_N\in U_N}\left(\left(S_N^{(1)} + S_N^{(2)}\right)u_N^2 - \right. \\ - 2S_N^{(1)}\left(\overline{y}_{N,N-1} + G_N - z_N\right)u_N + \\ + S_N^{(1)}\left(z_N^2 - 2z_N\left(\overline{y}_{N,N-1} + G_N\right) + \right. \\ \left.\left. + \mathbf{M}\left[\left(y_N + G_N\right)^2 | \mathfrak{J}_{N-1}^y\right]\right)\right),$$

откуда с учетом независимости последних двух слагаемых от u_N и положительности коэффициента при u_N^2 следует:

$$u_N^* = \frac{S_N^{(1)}\left(\overline{y}_{N,N-1} + G_N\right) - S_N^{(1)} z_{N-1}}{S_N^{(1)} + S_N^{(2)}} = \\ = \frac{1}{R(N)}\left(Q_0(N)\left(\overline{y}_{N,N-1} + G_N\right) - P(N)z_{N-1}\right).$$

Кроме того, получаем и выражение для функции Беллмана:

$$B(N) = L(N)z_{N-1}^2 - \\ -2Q_1(N-1)z_{N-1}\left(\overline{y}_{N,N-1} + G_N\right) + \mathbf{M}\left[A(N) | \mathfrak{J}_{N-1}^y\right],$$

где обозначено:

$$A(N) = S_N^{(1)}\left(y_N + G_N\right)^2 - \\ - \frac{1}{R(N)}\left(Q_0(N)\left(\overline{y}_{N,N-1} + G_N\right)\right)^2.$$

Предположим, что утверждение теоремы выполнено для u^*_{t+1} и для функции Беллмана $B(t+1)$ имеет место следующее выражение:

$$B(t+1) = L(t+1)z_t^2 - \\ - 2z_t \sum_{j=1}^{N-t} Q_j(t)\left(\overline{y}_{t+j,t} + G_{t+j}\right) + \mathrm{M}\left[A(t+1)|\mathfrak{J}_t^y\right],$$

где

$$A(t+1) = A(t+2) + S_{t+1}^{(1)}\left(y_{t+1} + G_{t+1}\right)^2 - \\ - \frac{1}{R(t+1)}\left(\sum_{j=1}^{N-t} Q_{j-1}(t+1)\left(\overline{y}_{t+j,t} + G_{t+j}\right)\right)^2.$$

Тогда уравнение Беллмана для $B(t)$ имеет следующий вид:

$$B(t) = \min_{u_t \in U_t} \mathbf{M}\left[S_t^{(1)}\left(z_t - G_t - y_t\right)^2 + S_t^{(2)}u_t^2 + \right. \\ + L(t+1)z_t^2 - 2z_t \sum_{j=1}^{N-t} Q_j(t)\left(\overline{y}_{t+j,t} + G_{t+j}\right) + \\ \left. + \mathbf{M}\left[A(t+1)\Big|\mathfrak{J}_t^y\right]\Big|\mathfrak{J}_{t-1}^y\right].$$

Преобразовав полученное уравнение с учетом $\mathfrak{J}_{t-1}^y$-измеримости функции u_t, формулы полного математического ожидания и равенств $\overline{y}_{t+j,t-1} = \mathbf{M}\left[\overline{y}_{t+j,t}|\mathfrak{J}_{t-1}^y\right] = \mathbf{M}\left[y_{t+j}|\mathfrak{J}_{t-1}^y\right]$ и $z_t = z_{t-1} + u_t$, запишем:

$$B(t) = \min_{u_t \in U_t}\left[\left(S_t^{(1)} + S_t^{(2)} + L(t+1)\right)u_t^2 - \right. \\ - 2\left(S_t^{(1)}\left(\overline{y}_{t,t-1} + G_t\right) + \right. \\ + \sum_{j=1}^{N-t} Q_j(t)\left(\overline{y}_{t+j,t-1} + G_{t+j}\right) - \\ \left.\left. - \left(S_t^{(1)} + L(t+1)\right)z_{t-1}\right)u_t + \right. \\ + \left(S_t^{(1)} + L(t+1)\right)z_{t-1}^2 - \\ - 2\left(S_t^{(1)}z_{t-1}\left(\overline{y}_{t,t-1} + G_t\right) + \right. \\ \left. + z_{t-1}\sum_{j=1}^{N-t} Q_j(t)\left(\overline{y}_{t+j,t-1} + G_{t+j}\right)\right) + \\ \left. + \mathbf{M}\left[S_t^{(1)}\left(y_t + G_t\right)^2 + A(t+1)\Big|\mathfrak{J}_{t-1}^y\right]\right].$$

Применяя в последнем выражении введенные в (8) обозначения, получаем:

$$B(t) = \min_{u_t \in U_t}\left[R(t)u_t^2 - \right. \\ - 2\left(\sum_{j=0}^{N-t} Q_j(t)\left(\overline{y}_{t+j,t-1} + G_{t+j}\right) - P(t)z_{t-1}\right)u_t + \\ + \left(S_t^{(1)} + L(t+1)\right)z_{t-1}^2 - \\ - 2z_{t-1}\sum_{j=0}^{N-t} Q_j(t)\left(\overline{y}_{t+j,t-1} + G_{t+j}\right) + \\ \left. + \mathbf{M}\left[S_t^{(1)}\left(y_t + G_t\right)^2 + A(t+1)\Big|\mathfrak{J}_{t-1}^y\right]\right].$$

Поскольку в полученном соотношении три последних слагаемых не зависят от u_t, то в предположении положительности $R(t)$ получается доказываемое соотношение (8) для u_t^*. Кроме того, подстановкой u_t^* подтверждается справедливость индуктивного предположения относительно функции Беллмана.

Для завершения доказательства остается показать, что $R(t) > 0$. Для этого достаточно показать, что $L(t+1) \geq 0$. Поскольку $L(N+1) = 0$, то требуемое неравенство следует из выражения

$$L(t+1) = \left(L(t+2) + S_{t+1}^{(1)}\right) - \frac{\left(L(t+1) + S_{t+1}^{(1)}\right)^2}{L(t+2) + S_{t+1}^{(1)} + S_{t+1}^{(2)}}$$

и неотрицательности $S_{t+1}^{(1)} + S_{t+1}^{(2)}$. Теорема доказана.

Замечание. В полученном утверждении используются оптимальные в среднем квадратическом прогнозы $\overline{y}_{t+j,t-1}$ наблюдений y_{t+j} по наблюдениям y_τ, $\tau \leq t-1$, $j = 0, 1, \ldots$ В теореме 2 работы [1] получены аналогичные прогнозы в случае $d = 0$. Используя их, нетрудно записать соответствующие соотношения для $\overline{y}_{t+j,t-1}$:

$$\overline{y}_{t+j,t-1} = d\overline{y}_{t+j-1,t-1} + \\ + \sum_{k=1}^{n}\int\limits_{\Delta_k} c(a_k\xi + q_k)\overline{\psi}_x(\xi,t,j)\,d\xi, \quad j = 1,2,\ldots;$$

$$\overline{y}_{t,t-1} = dy_{t-1} + \sum_{k=1}^{n}\int\limits_{\Delta_k} c(a_k\xi + q_k)\hat{\psi}_x(\xi,t-1)\,d\xi,$$

где прогнозирующие плотности вероятности определяются соотношениями:

$$\overline{\psi}_x(x,t,j) = \\ = \sum_{k=1}^{n}\frac{1}{b_k}\int\limits_{\Delta_k}\overline{\psi}_x(\xi,t,j-1)\varphi_v\left(\frac{x - a_k\xi - q_k}{b_k}\right)d\xi, \\ j = 1,2,\ldots;$$

$$\overline{\psi}_x(x,t,0) \triangleq$$
$$\triangleq \sum_{k=1}^{n} \frac{1}{b_k} \int\limits_{\Delta_k} \hat{\psi}_x(\xi,t-1)\varphi_v\left(\frac{x-a_k\xi-q_k}{b_k}\right) d\xi\,;$$
$$\hat{\psi}_x(x,t) = \left(\varphi_w\left(\frac{y_t - dy_{t-1} - cx}{\sigma}\right)\times\right.$$
$$\left.\times \sum_{k=1}^{n} \frac{1}{b_k} \int\limits_{\Delta_k} \hat{\psi}_x(\xi,t-1)\varphi_v\left(\frac{x-a_k\xi-q_k}{b_t^k}\right) d\xi\right) \Bigg/$$
$$\Bigg/ \left(\sum_{k=1}^{n} \frac{1}{b_t^k} \int\limits_{R^1} \varphi_w\left(\frac{y_t - dy_{t-1} - cx}{\sigma}\right)\times\right.$$
$$\left.\times \int\limits_{\Delta_k} \hat{\psi}_x(\xi,t-1)\varphi_v\left(\frac{x-a_k\xi-q_k}{b_k}\right) d\xi dx\right).$$

5 Субоптимальные управления

Основной проблемой применения оптимальной стратегии оптимизации пользовательского хранилища (8), как и определения размера пула в [2], является определение горизонта N. При этом надо учитывать, что выбор больших значений N приводит к необходимости выполнения вычислительно затратных расчетов большого числа прогнозов, а использование малых N лишает задачу характерных динамических свойств. Преодолевать указанную сложность предлагается, как и в [2], за счет практического применения субоптимальной стратегии оптимизации, основанной на принципе локально-оптимального (адаптивного) управления [7]. Выполняя локальную оптимизацию целевой функции (6), определим в качестве субоптимального решения функцию u_t^L, доставляющую минимум функционалу

$$\mathbf{J}_t(u_t) = \mathbf{M}\left[S_t^{(1)}\left(z_t - G_t - y_t\right)^2 + S_t^{(2)}u_t^2 + \right.$$
$$\left. + S_{t+1}^{(1)}\left(z_{t+1} - G_{t+1} - y_{t+1}\right)^2 + S_{t+1}^{(2)}u_{t+1}^2\right].$$

Таким образом, для локально-оптимального решения рассматриваемой задачи оптимизации предлагается двухшаговый вариант целевой функции вида (6), обновляемый на каждом следующем шаге алгоритма. В целевую функцию $\mathbf{J}_t(u_t)$, как легко видеть, включены штрафы за ошибку в определении размера пользовательского хранилища и за его размер на текущем и следующем шаге.

Требующееся выражение для функции $u_t^L = \arg\min_{u_t \in U_t} \mathbf{J}_t(u_t)$ получаем непосредственно как частный случай (8):

$$u_t^L =$$
$$= \frac{1}{S_t^{(1)} + S_t^{(2)} + S_{t+1}^{(1)} - \left(S_{t+1}^{(1)}\right)^2 \Big/ \left(S_{t+1}^{(1)} + S_{t+1}^{(2)}\right)} \times$$
$$\times \left(S_t^{(1)}\left(\overline{y}_{t,t-1} + G_t\right) + \left(S_{t+1}^{(1)} - \frac{\left(S_{t+1}^{(1)}\right)^2}{S_{t+1}^{(1)} + S_{t+1}^{(2)}}\right) \times \right.$$
$$\times \left(\overline{y}_{t+1,t-1} + G_{t+1}\right) -$$
$$\left. - \left(S_t^{(1)} + S_{t+1}^{(1)} - \frac{\left(S_{t+1}^{(1)}\right)^2}{S_{t+1}^{(1)} + S_{t+1}^{(2)}}\right) z_{t-1}\right), \quad (9)$$

Наконец, как и в [2], будем использовать самый простой вариант возможного решения рассматриваемой задачи оптимизации — оптимальную программную стратегию u_t^P в виде:

$$u_t^P = \mathbf{M}\left[u_t^*\right]. \quad (10)$$

6 Результаты численных экспериментов

Для сравнения предложенных алгоритмов оптимизации использован незначительно измененный модельный пример из [1]. Заданы три интервала $\Delta_1 = (-\infty;3]$, $\Delta_2 = (3;7]$, $\Delta_3 = (7;+\infty)$ и следующие параметры уравнения (1):

a_1	a_2	a_3	q_1	q_2	q_3	b_1	b_2	b_3
0,3	0,4	0,7	1,4	3,0	3,0	0,9	1,5	2,5

Параметры наблюдений (3): $c = 2{,}5$, $d = 0{,}5$, $\sigma = 1{,}0$. Распределения всех возмущений — стандартные гауссовские, распределение начального условия x_0 также предполагалось гауссовским со средним и дисперсией, равными соответствующим моментам предельного распределения (подробнее см. [1]).

Расчеты проводились для 10 шагов траектории: $t = 1,2,\ldots,10$, для вычисления значения целевой функции использовался пучок из 10 000 траекторий. Параметры целевой функции (6) выбраны следующим образом:

$$S_t^{(1)} = S_t^{(2)} = 0{,}1\,,\ \ 1 \le t \le 10\,;$$
$$G_0 = 0{,}0\,,\ \ G_1 = G_2 = G_3 = 0{,}1\,;$$
$$G_4 = G_5 = G_6 = 10{,}0\,;$$
$$G_7 = G_8 = G_9 = G_{10} = 40{,}0\,.$$

Для сравнения качества стратегий оптимизации (8)–(10) вычислялось значение целевой функции в каждый момент t, т. е. $\mathbf{J}(u_0,\ldots,u_t)$.

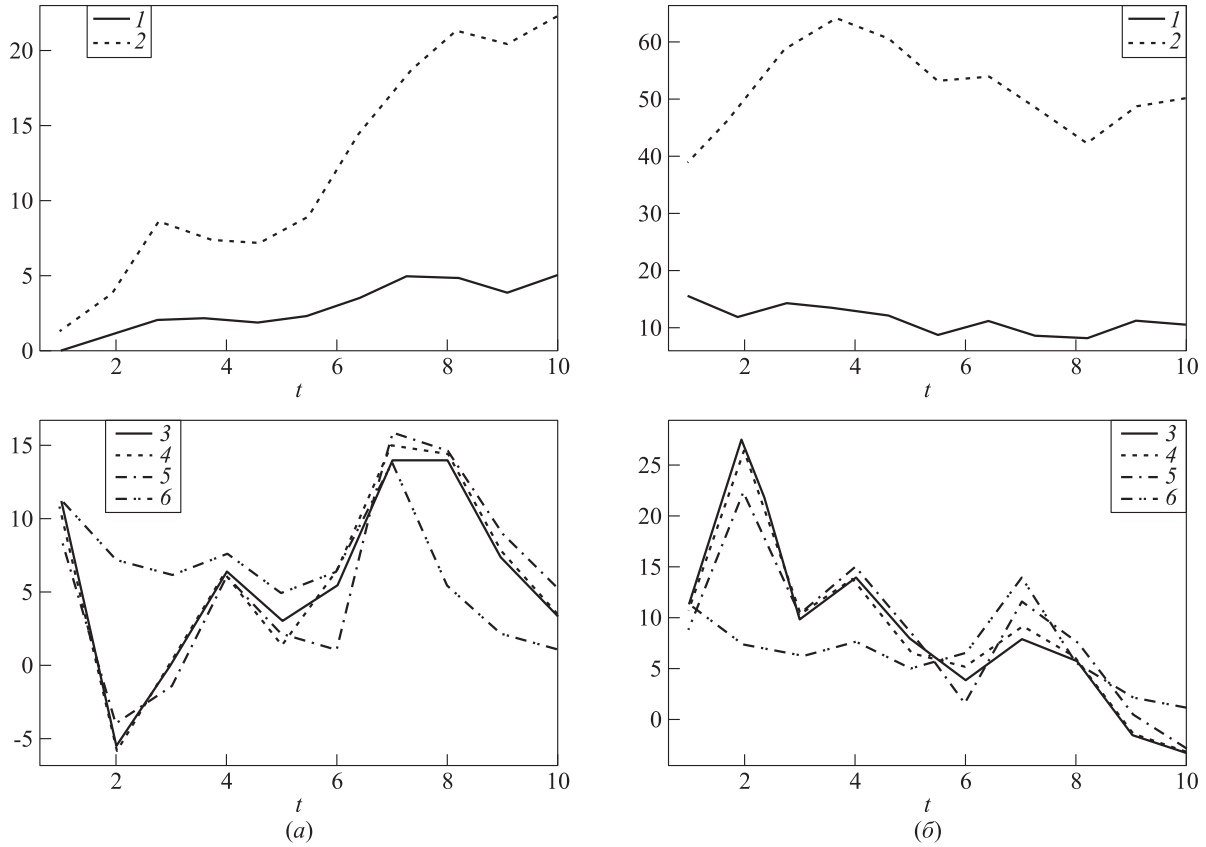


Рис. 1 Характерные траектории: *1* — x_t; *2* — y_t; *3* — u_t^*; *4* — u_t^{L1}; *5* — u_t^{L2}; *6* — u_t^P

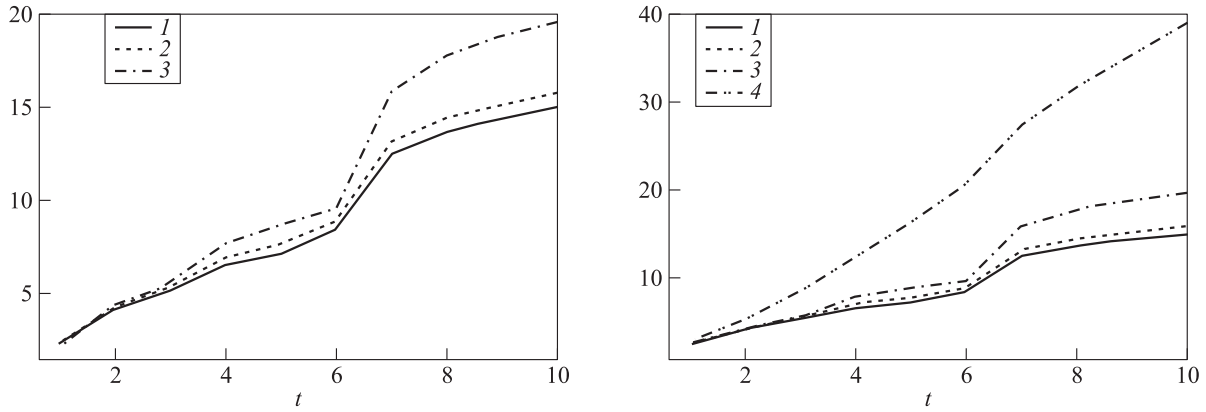


Рис. 2 Оценки целевых функций: *1* — $\hat{J}_1$; *2* — $\hat{J}_2$; *3* — $\hat{J}_3$; *4* — $\hat{J}_4$

Локально-оптимальная стратегия вычислялась двумя способами:

(*а*) функция u_t^{L1} минимизировала $\mathbf{J}_t(u_t)$, используя указанные выше значения $S_t^{(1)}$, $S_t^{(2)}$, $S_{t+1}^{(1)}$, $S_{t+1}^{(2)}$;

(*б*) функция u_t^{L2} минимизировала $\mathbf{J}_t(u_t)$ для таких же $S_t^{(1)}$, $S_t^{(2)}$ и $S_{t+1}^{(1)} = S_{t+1}^{(2)} = 0$.

Таким образом, стратегию u_t^{L1} можно назвать локально-оптимальной двухшаговой, стратегию u_t^{L2} — локально-оптимальной одношаговой.

Результаты расчетов приведены на рис. 1 и 2. Рисунок 1 иллюстрирует характерные траектории компонентов системы наблюдения (1)–(3) и функций u_t^*, u_t^{L1}, u_t^{L2}, u_t^P, на рис. 2 приведены значения целевых функций $\hat{J}_1$ (для u_t^*), $\hat{J}_2$ (для u_t^{L1}), $\hat{J}_3$ (для

u_t^{L2}) и $\hat{J}_4$ (для u_t^P). На рис. 2 приведены два графика: на первом изображены кривые для трех функций $\hat{J}_1$, $\hat{J}_2$ и $\hat{J}_3$, на втором — все четыре целевых функционала. Это сделано для того, чтобы позволить качественно оценить разницу оптимальной стратегии в сравнении с локально-оптимальными, которая становится несущественной в масштабе целевой функции $\hat{J}_4$.

Из приведенных результатов видно, что значения целевой функции, достигаемые при использовании оптимальной стратегии оптимизации u_t^* и локально-оптимальной двухшаговой u_t^{L1}, отличаются несущественно. Их преимущество в сравнении с локально-оптимальной одношаговой стратегией u_t^{L2} достигает 25%–30%. Преимущество же u_t^* в сравнении с u_t^{L1} составляет порядка 5%–7% к моменту достижения горизонта оптимизации N. При этом очевидно, что вычислительные трудности, возникающие в связи с необходимостью расчета большого числа прогнозов, весьма велики и можно считать, таким образом, что наиболее целесообразно использование локально-оптимальной двухшаговой стратегии оптимизации размера персонального пользовательского хранилища. Также нужно отметить, что программная стратегия u_t^P существенно проигрывает оптимальным.

7 Заключение

В статье завершено исследование модели пользовательской активности, предложенной в [1]. Стохастическая динамическая система наблюдения (1)–(3), описывающая эволюцию числа пользователей, взаимодействующих с некоторой информационной системой, и косвенные наблюдения за ними — объем персональной информации, размещенной в пользовательском хранилище (личном кабинете), дополнено уравнением выхода (4), описывающим размер базы данных, используемой для размещения пользовательской информации. В соответствии с предложенной в [2] терминологией используемый таким образом вычислительный ресурс назван «внешним», так как его обслуживанием занимается «внешняя» обслуживающая система — СУБД.

Основная решенная задача состоит в оптимизации в процессе работы программы размера пользовательской базы данных на основе квадратичного критерия качества, позволяющего учесть пользовательские потребности и издержки, определяемые размером базы данных. В целом предложенный подход аналогичен использованному в [2] и соответствует классической задаче динамического управления по квадратичному критерию качества. Рассмотренная постановка, однако, имеет существенное отличие: управляющее воздействие не влияет на фазовый процесс (текущее число пользователей), а входит аддитивно в уравнение наблюдений. Такой результат вполне согласуется с физическим смыслом рассматриваемой задачи — наилучшим образом распределять ресурсы «внешней» обслуживающей системы, состояние которой является одним из факторов состояния внешней среды.

Литература

1. *Босов А. В.* Задачи анализа и оптимизации для модели пользовательской активности. Часть 1. Анализ и прогнозирование // Информатика и её применения, 2011. Т. 5. Вып. 4. С. 40–52.
2. *Босов А. В.* Задачи анализа и оптимизации для модели пользовательской активности. Часть 2. Оптимизация внутренних ресурсов // Информатика и её применения, 2012. Т. 6. Вып. 1. С. 18–25.
3. Информационный веб-портал. Свидетельство об официальной регистрации программы для ЭВМ № 2005612992. Зарегистрировано в Реестре программ для ЭВМ 18.11.2005.
4. *Босов А. В.* Моделирование и оптимизация процессов функционирования Информационного web-портала // Программирование, 2009. № 6. С. 53–66.
5. *Флеминг У., Ришел Р.* Оптимальное управление детерминированными и стохастическими системами. — М.: Мир, 1978.
6. *Бертсекас Д., Шрив С.* Стохастическое оптимальное управление. — М.: Наука, 1985.
7. *Коган М. М., Неймарк Ю. И.* Адаптивное локально-оптимальное управление // Автоматика и телемеханика, 1987. № 8. С. 126–136.

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2012. Т. 6. Вып. 2. С. 22–28

ОБ УСТОЙЧИВОСТИ СДВИГОВЫХ СМЕСЕЙ НОРМАЛЬНЫХ ЗАКОНОВ ПО ОТНОШЕНИЮ К ИЗМЕНЕНИЯМ СМЕШИВАЮЩЕГО РАСПРЕДЕЛЕНИЯ

А. К. Горшенин[1]

Аннотация: Работа посвящена изучению устойчивости конечных сдвиговых смесей нормальных законов относительно изменений параметров смешивающего распределения. Результаты формулируются для моделей добавления и расщепления компоненты, которые используются в задачах проверки статистических гипотез о числе компонент смеси.

Ключевые слова: сдвиговые смеси нормальных законов; метрика Леви

1 Введение

В работе рассмотрены две модели конечных сдвиговых смесей нормальных распределений: добавления и расщепления компоненты (подробнее см. работы [1, 2]). Данные модели являются весьма удобными и информативными при проведении статистического анализа данных с помощью различных итерационных процедур разделения смесей вероятностных распределений и позволяют эффективно решать задачи проверки гипотез о числе компонент смеси.

Модель добавления компоненты удобно использовать для проверки значимости компоненты с малым весом. Дело в том, что при статистическом определении параметров в модели типа смеси вероятностных распределений может появиться компонента, вес которой значительно меньше весов остальных компонент. В такой ситуации необходимо убедиться в статистической значимости этой компоненты, чтобы избежать влияния погрешностей вычисления на итоговый результат.

Модель расщепления компоненты может применяться для в некотором смысле обратной задачи, когда в силу вычислительных ошибок компонента с малым весом может быть ошибочно отнесена к одной из компонент с бóльшим весом. Наличие устойчивости играет важную роль при практическом использовании данных моделей, так как гарантирует корректность полученных результатов.

В работе [3] были получены результаты для конечных масштабных смесей нормальных законов. Однако в ряде ситуаций оказывается полезным рассмотрение сдвиговых конечных нормальных смесей. Модели такого типа возникают, например, при решении оптимизационных задач для управлении запасами, при моделировании потоков страховых выплат, при прогнозировании надежности различных систем. Доказательству теорем устойчивости для сдвиговых конечных смесей нормальных законов и посвящена настоящая статья.

2 Постановка задачи

Предположим, что каждое из независимых наблюдений $\mathbf{X}_n = (X_1, \ldots, X_n)$ имеет распределение, представимое в виде конечной сдвиговой смеси нормальных законов, т. е.

$$G(x) = \sum_{i=1}^{k} p_i \Phi(x - a_i)\,, \tag{1}$$

где

$$\sum_{i=1}^{k} p_i = 1\,, \quad p_i \geqslant 0\,, \;\; a_i \in \mathbb{R}\,, \;\; i = 1, \ldots, k,$$

а через $\Phi(\cdot)$ обозначена функция распределения стандартного нормального закона

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} \exp\left\{-\frac{t^2}{2}\right\} dt\,.$$

Также в дальнейшем будет использоваться функция плотности стандартного нормального закона

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}.$$

*Работа выполнена при поддержке Российского фонда фундаментальных исследований (проекты 11-01-12026-офи-м и 12-07-00115).

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

Очевидно, что функция распределения $G(x)$ из соотношения (1) может быть представлена в виде:

$$G(x) = \mathbb{E}\Phi(x - V)\,,$$

где V — дискретная случайная величина, принимающая значения a_i с вероятностями p_i, т. е.

$$V: \begin{matrix} a_1 & a_2 & \cdots & a_k \\ p_1 & p_2 & \cdots & p_k\,. \end{matrix} \qquad (2)$$

Обозначим через $\rho(F, G)$ равномерное расстояние между функциями распределения $F(x)$ и $G(x)$:

$$\rho(F, G) = \sup_{x\in\mathbb{R}} |F(x) - G(x)|\,. \qquad (3)$$

Известно, что для решения задачи устойчивости для конечных сдвиговых смесей нормальных законов равномерная метрика (3) является не вполне корректной (можно привести пример весьма близких функций распределения, для которых равномерная метрика будет давать расстояние, равное единице). Поэтому необходимо рассматривать метрики, метризующие слабую сходимость, например метрику Леви $L(F, G)$ между функциями распределения $F(x)$ и $G(x)$, определяемую соотношением:

$$L(F, G) = \inf\{h:\ G(x - h) - h \leqslant \\ \leqslant F(x) \leqslant G(x + h) + h, \forall x \in \mathbb{R}\}\ .$$

Модели добавления и расщепления компоненты могут быть представлены в виде:

$$G_p(x) = \mathbb{E}\Phi(x - V_p)\,,$$

где дискретная случайная величина V_p определяется для каждой из моделей по-разному. Необходимо получить соотношения, связывающие расстояния Леви между смешивающими распределениями и смесями. Перейдем к рассмотрению каждой из моделей.

3 Модель добавления компоненты

Модель добавления компоненты формализуется следующим образом. Предполагается, что каждое из независимых наблюдений $\mathbf{X}_n = (X_1, \ldots, X_n)$ имеет распределение, представимое в виде:

$$G_p(x) = (1 - p)\sum_{i=1}^{k} p_i\Phi(x - a_i) + p\Phi(x - a)\,, \qquad (4)$$

где все величины $a_i \in \mathbb{R}$, $p_i \geqslant 0$, $i = 1, \ldots, k$, считаются известными, а a и p являются параметрами модели, при этом $a \in \mathbb{R}$, $0 \leqslant p \leqslant 1$. Без ограничения общности для определенности будем считать, что выполнены соотношения

$$a_0 \leqslant a \leqslant a_1 \leqslant a_2 \leqslant \cdots \leqslant a_k\,. \qquad (5)$$

Левое неравенство означает достаточно естественное для практики предположение, что рассматриваются конечные математические ожидания. Поэтому в дальнейшем считаем a_0 известным параметром модели (так как он может быть указан из некоторых разумных предположений для каждого конкретного случая).

В модели добавления компоненты дискретная случайная величина V_p имеет следующий вид:

$$V_p: \begin{matrix} a & a_1 & a_2 & \cdots & a_k \\ p & p_1(1-p) & p_2(1-p) & \cdots & p_k(1-p). \end{matrix} \qquad (6)$$

Отметим, что расстояние Леви $L(V, V_p)$ не превосходит величины p, так как расстояние между ступеньками функций распределения составляет в точности p на сегменте $[a, a_1]$ и pp_i на сегментах $[a_i, a_{i+1}]$, $i = 1, \ldots, k - 1$. Изменяться могут лишь параметры a и p, величины a_i, p_i, $i = 1, \ldots, k$, считаем постоянными. Однако при фиксированном параметре p и при $a \to a_1$ очевидно, что $L(V, V_p)$ к нулю не стремится. Таким образом, без ограничения общности считаем, что $0 \leqslant p \leqslant a_1 - a$. Поэтому

$$L(V, V_p) = p\,. \qquad (7)$$

Тогда справедлива следующая теорема.

Теорема 1. *В рамках модели добавления компоненты* (4) *при выполнении условий* (5) *и* (7) *расстояние Леви $L(V, V_p)$ между смешивающими распределениями V из соотношения* (2) *и V_p из соотношения* (6) *и расстояние Леви $L(G, G_p)$ между истинным распределением $G(x)$ из соотношения* (1) *и приближающей смесью $G_p(x)$ из соотношения* (4) *связывают неравенства*

$$C_1^{[1]}(a_0, a_k)L(G, G_p) \leqslant L(V, V_p) \leqslant \\ \leqslant C_2^{[1]}(a_0, a_k)L^{1/2}(G, G_p)\,,$$

где коэффициенты $C_j^{[1]}(a_0, a_k)$, $j = 1, 2$, зависящие только от известных величин a_k и a_0, имеют вид:

$$C_1^{[1]}(a_0, a_k) = \max\left\{1, \frac{\sqrt{2\pi}}{a_k - \min\{0, a_0\}}\right\}\,; \qquad (8)$$

$$C_2^{[1]}(a_0, a_k) = \varphi^{-1/2}\left(a_k + |a_k| - \right. \\ \left. - \min\{0, a_0\}\right)\left(1 + \frac{1}{\sqrt{2\pi}}\right)^{1/2}\,. \qquad (9)$$

Доказательство. Запишем оценки снизу для равномерного расстояния между функциями распределения $G(x)$ и $G_p(x)$, воспользовавшись формулой Лагранжа:

$$\rho(G,G_p)=\sup_x|G(x)-G_p(x)|=$$
$$=\sup_x|G(x)-G(x)+p(G(x)-\Phi(x-a))|=$$
$$=p\sup_x|G(x)-\Phi(x-a)|\geqslant p|G(x_0-a_i)-\Phi(x_0-a)|=$$
$$=p\left|\sum_{i=1}^k p_i\left(\Phi(x_0-a_i)-\Phi(x_0-a)\right)\right|=$$
$$=p\left|\sum_{i=1}^k p_i(a-a_i)\varphi(\theta_i(x_0-a_i)+(1-\theta_i)(x_0-a))\right|=$$
$$=p\left|\sum_{i=1}^k p_i(a_i-a)\varphi(x_0-a-\theta_i(a_i-a))\right|. \quad (10)$$

Неравенство в соотношении (10) справедливо для любого x_0. Выберем значение данной величины так, чтобы воспользоваться свойством монотонного убывания плотности стандартного нормального распределения $\varphi(x)$ от положительного аргумента. А именно потребуем выполнения условия

$$x_0-a-\theta_i(a_i-a)\geqslant 0\,,$$

откуда следует (с учетом того, что выражение в скобках в силу условий (5) неотрицательно и $0\leqslant\theta_i\leqslant 1$), что

$$x_0\geqslant a_i \quad (11)$$

сразу для всех номеров i. Тогда в качестве x_0 возьмем величину

$$x_0=a_k+|a_k|\,. \quad (12)$$

Очевидно, что условие (11) выполняется, при этом $x_0\geqslant 0$ и $x_0-a\geqslant 0$. Тогда, продолжая (10) с учетом соотношений (5) и (7), получим:

$$\rho(G,G_p)\geqslant$$
$$\geqslant p\left|\sum_{i=1}^k p_i(a_i-a)\varphi\left(a_k+|a_k|-a-\theta_i(a_i-a)\right)\right|\geqslant$$
$$\geqslant p\left|\sum_{i=1}^k p_i(a_i-a)\varphi\left(a_k+|a_k|-a\right)\right|\geqslant$$
$$\geqslant p\left|\sum_{i=1}^k p_i(a_i-a)\varphi\Big(a_k+|a_k|-\min\{0,a_0\}\Big)\right|\geqslant$$
$$\geqslant p\sum_{i=1}^k p_i(a_1-a)\varphi\left(a_k+|a_k|-\min\{0,a_0\}\right)=$$
$$=p(a_1-a)\varphi\left(a_k+|a_k|-\min\{0,a_0\}\right)\geqslant$$
$$\geqslant L^2(V,V_p)\varphi\left(a_k+|a_k|-\min\{0,a_0\}\right).$$

Воспользуемся известным неравенством для метрики Леви (см., например, книгу [4]):

$$L(G,G_p)\leqslant\rho(G,G_p)\leqslant$$
$$\leqslant(1+\max_x G'(x))L(G,G_p)\,. \quad (13)$$

Воспользуемся правым неравенством из соотношения (13). Имеем

$$L^2(V,V_p)\varphi\left(a_k+|a_k|-\min\{0,a_0\}\right)\leqslant\rho(G,G_p)\leqslant$$
$$\leqslant(1+\max_x G'(x))L(G,G_p)=$$
$$=\left(1+\max_x\left(\sum_{i=1}^k p_i\varphi(x-a_i)\right)\right)L(G,G_p)\leqslant$$
$$\leqslant\left(1+\sum_{i=1}^k p_i\frac{1}{\sqrt{2\pi}}\right)L(G,G_p)=$$
$$=\left(1+\frac{1}{\sqrt{2\pi}}\right)L(G,G_p)\,.$$

Окончательно получаем следующую оценку сверху для $L(V,V_p)$:

$$L(V,V_p)\leqslant\varphi^{-1/2}\left(a_k+|a_k|-\min\{0,a_0\}\right)\times$$
$$\times\left(1+\frac{1}{\sqrt{2\pi}}\right)^{1/2}L^{1/2}(G,G_p)=$$
$$=C_2^{[1]}(a_0,a_k)L^{1/2}(G,G_p)\,.$$

Оценка снизу для $L(V,V_p)$ может быть найдена из соотношений

$$L(G,G_p)\leqslant\rho(G,G_p)=\sup_x|G(x)-G_p(x)|=$$
$$=p\sup_x\left|\sum_{i=1}^k p_i(\Phi(x-a_i)-\Phi(x-a))\right|\leqslant$$
$$\leqslant p\sup_x\sum_{i=1}^k p_i|\Phi(x-a_i)-\Phi(x-a)|\leqslant$$
$$\leqslant p\sum_{i=1}^k p_i\sup_x|\Phi(x-a_i)-\Phi(x-a)|\leqslant$$
$$\leqslant p\sum_{i=1}^k p_i=L(V,V_p)\,.$$

Однако можно провести оценивание и другим путем. Найдем точки экстремума функции $\Phi(x-a)-\Phi(x-a_i)$ из условия

$$\varphi(x-a)-\varphi(x-a_i)=0\,.$$

Максимум достигается в точке

$$x_i^*=\frac{a+a_i}{2}\,.$$

Тогда, учитывая четность функции $\varphi(x)$, получим:

$$p\sum_{i=1}^{k} p_i \sup_x |\Phi(x-a_i)-\Phi(x-a)| \leqslant$$
$$\leqslant p\sup_x \sum_{i=1}^{k} p_i |\Phi(x-a_i)-\Phi(x-a)| \leqslant$$
$$\leqslant p\sup_x \sum_{i=1}^{k} p_i |\Phi(x_i^*-a_i)-\Phi(x_i^*-a)| =$$
$$= p\sum_{i=1}^{k} p_i (a_i-a)\varphi\left(\theta(x_i^*-a)+(1-\theta)(x_i^*-a_i)\right) \leqslant$$
$$\leqslant p\frac{a_k-\min\{0,a_0\}}{\sqrt{2\pi}} = L(V,V_p)\frac{a_k-\min\{0,a_0\}}{\sqrt{2\pi}}\,.$$

Окончательно

$$L(V,V_p) \geqslant \max\left\{1, \frac{\sqrt{2\pi}}{a_k-\min\{0,a_0\}}\right\} L(G,G_p) =$$
$$= C_1^{[1]}(a_0,a_k)L(G,G_p)\,. \;\square$$

Рассмотрим следующее обобщение модели (4). Пусть имеется еще одна смесь данного типа, отличающаяся от (4) только весом, т. е. (при этом $0 \leqslant q \leqslant 1$)

$$G_q(x) = (1-q)\sum_{i=1}^{k} p_i\Phi(x-a_i) + q\Phi(x-a)\,. \quad (14)$$

Для $G_q(x)$ дискретная случайная величина V_q имеет следующий вид:

$$V_q: \begin{array}{ccccc} a & a_1 & a_2 & \cdots & a_k \\ q & p_1(1-q) & p_2(1-q) & \cdots & p_k(1-q) \end{array}. \quad (15)$$

Рассуждая как описано выше, получим, что $|p-q| \leqslant a_1 - a$. В этом случае расстояние Леви $L(V_p,V_q)$ примет вид:

$$L(V_p,V_q) = |p-q|\,. \quad (16)$$

Тогда справедлива следующая теорема.

Теорема 2. *В рамках модели добавления компоненты* (4) *при выполнении условий* (5) *и* (16) *расстояние Леви* $L(V_p,V_q)$ *между смешивающими распределениями* V_p *из соотношения* (6) *и* V_q *из соотношения* (15) *и расстояние Леви* $L(G_p,G_q)$ *между распределениями* $G_p(x)$ *из соотношения* (4) *и* $G_q(x)$ *из соотношения* (14) *связывают неравенства:*

$$C_1^{[1]}(a_0,a_k)L(G_p,G_q) \leqslant L(V_p,V_q) \leqslant$$
$$\leqslant C_2^{[1]}(a_0,a_k)L^{1/2}(G_p,G_q)\,,$$

где коэффициенты $C_j^{[1]}(a_0,a_k)$, $j=1,2$, *зависящие только от известных величин* a_k *и* a_0, *определяются формулами* (8) и (9).

Доказательство. Рассуждая аналогично доказательству теоремы 1, найдем оценки снизу для равномерного расстояния между функциями распределения $G_p(x)$ и $G_q(x)$. Имеем:

$$\rho(G_p,G_q) =$$
$$= \sup_x |(q-p)\sum_{i=1}^{k} p_i\Phi(x-a_i) + (p-q)\Phi(x-a)| =$$
$$= |p-q|\sup_x \left|\sum_{i=1}^{k} p_i\Phi(x-a_i) - \Phi(x-a)\right| \geqslant$$
$$\geqslant |p-q|\left|\sum_{i=1}^{k} p_i(\Phi(x-a_i)-\Phi(x-a))\right| \geqslant$$
$$\geqslant L^2(V_p,V_q)\varphi\left(a_k+|a_k|-\min\{0,a_0\}\right)\,.$$

Оценим максимум производной для функций G_p и G_q. Запишем выражения, например, для функции G_p (для функции G_q оценка получается аналогично). Имеем:

$$\max_x G_p'(x) =$$
$$= \max_x \left((1-p)\sum_{i=1}^{k} p_i\varphi(x-a_i) + p\varphi(x-a)\right) \leqslant$$
$$\leqslant \frac{1-p}{\sqrt{2\pi}}\sum_{i=1}^{k} p_i + \frac{p}{\sqrt{2\pi}} = \frac{1}{\sqrt{2\pi}}\,.$$

Пользуясь правым неравенством в формуле (13), приходим к следующему результату:

$$L(V_p,V_q) \leqslant \varphi^{-1/2}\left(a_k+|a_k|-\min\{0,a_0\}\right)\times$$
$$\times\left(1+\frac{1}{\sqrt{2\pi}}\right)^{1/2} L^{1/2}(G_p,G_q) =$$
$$= C_2^{[1]}(a_0,a_k)L^{1/2}(G_p,G_q)\,.$$

Оценка снизу для $L(V_p,V_q)$ может быть найдена из следующих соотношений:

$$L(G_p,G_q) \leqslant \rho(G_p,G_q) =$$
$$= |p-q|\sup_x \left|\sum_{i=1}^{k} p_i\Phi(x-a_i) - \Phi(x-a)\right| \leqslant$$
$$\leqslant |p-q|\sup_x \sum_{i=1}^{k} p_i|\Phi(x-a_i)-\Phi(x-a)| \leqslant$$

$$\leqslant |p-q| \sum_{i=1}^{k} p_i \sup_x |\Phi(x-a_i) - \Phi(x-a)| \leqslant$$
$$\leqslant |p-q| \sum_{i=1}^{k} p_i = L(V_p, V_q)\,.$$

Аналогично доказательству теоремы 1 получим:

$$L(V_p, V_q) \geqslant \max\left\{1, \frac{\sqrt{2\pi}}{a_k - \min\{0, a_0\}}\right\} L(G_p, G_q) =$$
$$= C_1^{[1]}(a_0, a_k) L(G_p, G_q)\,. \;\square$$

4 Модель расщепления компоненты

Модель расщепления компоненты формализуется следующим образом. Предполагается, что каждое из независимых наблюдений $\mathbf{X}_n = (X_1, \ldots, X_n)$ имеет распределение, представимое в виде:

$$G_p(x) = \sum_{i=1}^{k-1} p_i \Phi(x-a_i) + (p_k - p)\Phi(x-a_k) + p\Phi(x-a)\,, \quad (17)$$

где все величины $a_i \in \mathbb{R}$, $0 \leqslant p_i \leqslant 1$, $i = 1, \ldots, k$, считаются известными, a и p являются параметрами модели, при этом $0 \leqslant p \leqslant p_k$. Без ограничения общности для определенности будем считать, что выполнены соотношения:

$$a_1 \leqslant a_2 \leqslant \cdots \leqslant a_{k-1} \leqslant a \leqslant a_k\,. \quad (18)$$

Для данной модели дискретная случайная величина V_p имеет вид:

$$V_p : \begin{array}{cccccc} a_1 & a_2 & \cdots & a & a_k \\ p_1 & p_2 & \cdots & p & p_k - p\,. \end{array} \quad (19)$$

Воспользовавшись геометрической интерпретацией расстояния Леви, можно получить, что

$$L(V, V_p) = \min\{a_k - a, p\}\,. \quad (20)$$

В этой ситуации оба условия: $a \to a_k$ при фиксированном параметре p и $p \to 0$ при фиксированном a — влекут справедливость соотношения $L(V, V_p) \to 0$. Тогда справедлива следующая теорема.

Теорема 3. *В рамках модели расщепления компоненты* (17) *при выполнении условий* (18) *расстояние Леви* $L(V, V_p)$ *из соотношения* (20) *между смешивающими распределениями* V *из соотношения* (2) *и* V_p *из соотношения* (19) *и расстояние Леви* $L(G, G_p)$ *между истинным распределением* $G(x)$ *из соотношения* (1) *и приближающей смесью* $G_p(x)$ *из соотношения* (17) *связывают неравенства:*

$$C_1^{[2]}(a_{k-1}, a_k) L(G, G_p) \leqslant L(V, V_p) \leqslant C_2^{[2]}(a_{k-1}, a_k) L^{1/2}(G, G_p)\,,$$

где коэффициенты $C_j^{[2]}(a_{k-1}, a_k)$, $j = 1, 2$, *не зависят от величин* a, p *и имеют вид:*

$$C_1^{[2]}(a_{k-1}, a_k) = \frac{\sqrt{2\pi}}{\max\{1, a_k - a_{k-1}\}}\,, \quad (21)$$
$$C_2^{[2]}(a_{k-1}, a_k) = \varphi^{-1/2}\left(a_k + |a_k| - \min\{0, a_{k-1}\}\right) \times \left(1 + \frac{1}{\sqrt{2\pi}}\right)^{1/2}\,. \quad (22)$$

Доказательство. Запишем оценки снизу для равномерного расстояния между функциями распределения $G(x)$ и $G_p(x)$, воспользовавшись формулой Лагранжа, свойством монотонного убывания плотности стандартного нормального распределения $\varphi(x)$ от положительного аргумента и соотношениями (12), (18) и (20):

$$\rho(G, G_p) = \sup_x |G(x) - G_p(x)| =$$
$$= \sup_x \left| \sum_{i=1}^{k} p_i \Phi(x-a_i) - \sum_{i=1}^{k} p_i \Phi(x-a_i) + p\Phi(x-a_k) - p\Phi(x-a) \right| =$$
$$= p \sup_x |\Phi(x-a_k) - \Phi(x-a)| \geqslant$$
$$\geqslant p|\Phi(x_0 - a_k) - \Phi(x_0 - a)| =$$
$$= p|(a_k - a)\varphi(\theta(x_0 - a_k) + (1-\theta)(x_0 - a))| \geqslant$$
$$\geqslant p(a_k - a)\varphi\left(a_k + |a_k| - \min\{0, a_{k-1}\}\right) \geqslant$$
$$\geqslant L^2(V, V_p)\varphi\left(a_k + |a_k| - \min\{0, a_{k-1}\}\right)\,.$$

Чтобы оценить сверху $L(V, V_p)$, воспользуемся правым неравенством из соотношения (13) и найденной в доказательстве теоремы 1 оценкой для максимума производной, а также неравенствами (18). Имеем:

$$L^2(V, V_p)\varphi\left(a_k + |a_k| - \min\{0, a_{k-1}\}\right) \leqslant \left(1 + \frac{1}{\sqrt{2\pi}}\right) L(G, G_p)\,,$$

откуда

$$L(V,V_p) \leqslant \varphi^{-1/2}\left(a_k + |a_k| - \min\{0, a_{k-1}\}\right) \times \times \left(1 + \frac{1}{\sqrt{2\pi}}\right)^{1/2} L^{1/2}(G,G_p) = = C_2^{[2]}(a_{k-1},a_k) L^{1/2}(G,G_p)\,.$$

Выпишем оценку снизу для $L(V,V_p)$. С этой целью заметим, что

$$L(G,G_p) \leqslant \rho(G,G_p) = p \sup_x |\Phi(x-a_k) - \Phi(x-a)| = = p \sup_x \left(\Phi(x-a) - \Phi(x-a_k)\right)\,. \quad (23)$$

Найдем точки экстремума функции $\Phi(x-a) - \Phi(x-a_k)$ из условия

$$\varphi(x-a) - \varphi(x-a_k) = 0\,.$$

Максимум достигается в точке:

$$x^* = \frac{a + a_k}{2}\,.$$

Подставляя это значение в (23), получим (учитывая четность функции $\varphi(x)$)

$$p \sup_x \left(\Phi(x-a) - \Phi(x-a_k)\right) = = p\left(\Phi(x^*-a) - \Phi(x^*-a_k)\right) = = p(a_k-a)\varphi\left(\theta(x^*-a) + (1-\theta)(x^*-a_k)\right) = = p(a_k-a)\varphi\left((a_k-a)\left|\theta - \frac{1}{2}\right|\right) \leqslant \leqslant L(V,V_p)\max\{p, a_k-a\}\frac{1}{\sqrt{2\pi}} \leqslant \leqslant L(V,V_p)\max\{1, a_k - a_{k-1}\}\frac{1}{\sqrt{2\pi}}\,.$$

Окончательно

$$L(V,V_p) \geqslant \frac{\sqrt{2\pi}}{\max\{1, a_k - a_{k-1}\}} L(G,G_p) = = C_1^{[2]}(a_{k-1},a_k) L(G,G_p)\,. \square$$

Рассмотрим следующее обобщение модели (17). Пусть имеется еще одна смесь данного типа, отличающаяся от (17) только весом, т. е. (при этом $0 \leqslant q \leqslant p_k$)

$$G_q(x) = \sum_{i=1}^{k-1} p_i \Phi(x-a_i) + + (p_k - q)\Phi(x-a_k) + q\Phi(x-a)\,. \quad (24)$$

Для $G_q(x)$ дискретная случайная величина V_q имеет вид

$$V_q: \begin{array}{ccccc} a_1 & a_2 & \cdots & a & a_k \\ p_1 & p_2 & \cdots & q & p_k - q \end{array}\,. \quad (25)$$

Воспользовавшись геометрической интерпретацией расстояния Леви, можно получить, что

$$L(V_p,V_q) = \min\{a_k - a, |p-q|\}\,. \quad (26)$$

Тогда справедлива следующая теорема.

Теорема 4. *В рамках модели расщепления компоненты* (17) *при выполнении условий* (18) *расстояние Леви* $L(V_p,V_q)$ *из соотношения* (26) *между смешивающими распределениями* V_p *из соотношения* (19) *и* V_q *из соотношения* (25) *и расстояние Леви* $L(G_p,G_q)$ *между распределениями* $G_p(x)$ *из соотношения* (17) *и* $G_q(x)$ *из соотношения* (24) *связывают неравенства:*

$$C_1^{[2]}(a_{k-1},a_k) L(G_p,G_q) \leqslant L(V_p,V_q) \leqslant \leqslant C_2^{[2]}(a_{k-1},a_k) L^{1/2}(G_p,G_q)\,,$$

где коэффициенты $C_j^{[2]}(a_{k-1},a_k)$, $j = 1, 2$, *не зависят от величин* a, p *и определяются формулами* (21) *и* (22).

Доказательство. Рассуждая аналогично доказательству теоремы 3, найдем оценки снизу для равномерного расстояния между функциями распределения $G_p(x)$ и $G_q(x)$. Имеем:

$$\rho(G_p,G_q) = \sup_x |G_p(x) - G_q(x)| = = |p-q| \sup_x |\Phi(x-a_k) - \Phi(x-a)| \geqslant \geqslant p|\Phi(x_0-a_k) - \Phi(x_0-a)| \geqslant \geqslant L^2(V_p,V_q)\varphi\left(a_k + |a_k| - \min\{0, a_{k-1}\}\right)\,.$$

Оценим максимум производной для функций G_p и G_q. Имеем:

$$\max_x G_p'(x) = \max_x \left(\sum_{i=1}^{k-1} p_i \varphi(x-a_i) + + (p_k - p)\varphi(x-a_k) + p\varphi(x-a)\right) \leqslant \leqslant \frac{1}{\sqrt{2\pi}} \sum_{i=1}^{k-1} p_i + \frac{(p_k-p)}{\sqrt{2\pi}} + \frac{p}{\sqrt{2\pi}} = \frac{1}{\sqrt{2\pi}}\,.$$

Пользуясь правым неравенством в формуле (13), приходим к следующему результату:

$$L(V_p,V_q) \leqslant \varphi^{-1/2}\left(a_k + |a_k| - \min\{0, a_{k-1}\}\right) \times \times \left(1 + \frac{1}{\sqrt{2\pi}}\right)^{1/2} L^{1/2}(G_p,G_q) = = C_2^{[2]}(a_{k-1},a_k) L^{1/2}(G_p,G_q)\,.$$

Оценка снизу для $L(V_p, V_q)$ может быть найдена из следующих соотношений:

$$L(G_p, G_q) \leqslant \rho(G_p, G_q) = \\ = |p-q| \sup_x |\Phi(x-a_k) - \Phi(x-a)| \,.$$

Повторяя рассуждения из доказательства теоремы 3, получаем, что

$$L(V_p, V_q) \geqslant \frac{\sqrt{2\pi}}{\max\{1, a_k - a_{k-1}\}} L(G, G_p) = \\ = C_1^{[2]}(a_{k-1}, a_k) L(G_p, G_q) \,. \ \square$$

5 Выводы

В рамках двух рассмотренных моделей возмущений параметров смеси — моделей добавления и расщепления компоненты — получены оценки устойчивости смесей нормальных законов по отношению к изменениям смешивающего параметра. Для каждой из моделей получены двусторонние оценки, связывающие расстояния Леви между смесями и смешивающими законами. Данные оценки, в частности, являются количественными характеристиками идентифицируемости конечных сдвиговых смесей нормальных законов.

В то же время доказанные теоремы 1–4 устанавливают взаимно однозначное соответствие между значением параметра веса и числом компонент в смеси. Данный результат удобно использовать при построении асимптотически наиболее мощных критериев для моделей добавления и расщепления компоненты для случая произвольных конечных сдвиг-масштабных смесей в качестве обоснования вида гипотез в задаче статистической проверки числа компонент смеси вероятностных распределений (подробнее об этом см. в работах [1, 5]).

Литература

1. *Бенинг В. Е., Горшенин А. К., Королев В. Ю.* Асимптотически оптимальный критерий проверки гипотез о числе компонент смеси вероятностных распределений // Информатика и её применения, 2011. Т. 5. Вып. 3. С. 4–16.
2. *Горшенин А. К.* Проверка статистических гипотез в модели расщепления компоненты // Вестник Московского университета. Сер. 15. Вычисл. матем. и киберн., 2011. № 4. С. 26–32.
3. *Горшенин А. К.* Устойчивость масштабных смесей нормальных законов по отношению к изменениям смешивающего распределения // Системы и средства информатики, 2012. Т. 22. Вып. 1. С. 136–148.
4. *Золотарев В. М.* Современная теория суммирования независимых случайных величин. — М.: Наука, 1986. 417 с.
5. *Gorshenin A. K.* Testing of statistical hypotheses in the splitting component model // Moscow University Computational Mathematics and Cybernetics, 2011. Vol. 35. No. 4. P. 176–183.

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2012. Т. 6. Вып. 2. С. 29–42

ОБРАБОТКА ГЕОПРОСТРАНСТВЕННОЙ ИНФОРМАЦИИ НА БАЗЕ РЕПОЗИТОРИЯ ГЕОИНФОРМАЦИОННОЙ СИСТЕМЫ

С. К. Дулин[1], И. Н. Розенберг[2], В. И. Уманский[3]

Аннотация: Эффективная интеграция данных, описывающих взаимодействующие компоненты, — ключ к успешному управлению всей системой функционирующих объектов. Нехватка интеграции данных может привести к существенной неэффективности оперативных, тактических и долгосрочных стратегий управления. Интегрированные системы управления могут помочь преодолеть эту неэффективность и улучшить координацию и рентабельность решений. Интеграция данных — безусловно, самое ответственное мероприятие для осуществления успешных стратегий управления. В работе описан подход к использованию централизованного репозитория корпоративных данных, позволяющего объединить пространственные и непространственные данные описания объектов. Репозиторий представляется средой для совместного использования и объединения данных и используемых программных продуктов.

Ключевые слова: управление объектами; интеграция данных; ГИС; репозиторий

1 Введение

Информационное обеспечение системы функционирования объектов должно учитывать много сложных, взаимозависимых и перегруженных процессов.

Успешная реализация стратегий управления в значительной степени определяется: совместным использованием и управлением данными жизненного цикла объектов и способностью поддерживать и координировать процессы многофункциональной работы на тактическом и стратегическом уровнях. Недостаток интеграции данных при управлении объектами может привести к существенной неэффективности. Интегрированные системы управления объектами могут преодолеть эту неэффективность и улучшить координацию и рентабельность решений управления [1].

Данные описания объектов характеризуются большим объемом, сложностью, взаимозависимостями и динамизмом. Эти данные могут существовать в несопоставимых форматах из-за наличия разнообразных источников данных и программных систем. Объединение этих данных в согласованную и унифицированную форму считается критическим мероприятием для успешного управления [2, 3].

Большинство разработанных в последние десятилетия инструментальных средств функционируют как автономные системы с ограниченной возможностью совместного использования информации с другими системами. Быстрое увеличение числа этих инструментальных средств создало так называемые «острова информации» и появление противоречивых моделей данных в несопоставимых программных продуктах.

Эффективная интеграция данных может значительно улучшить рентабельность принятия решений при управлении объектами на оперативном, тактическом и стратегическом уровнях. В результате интеграции данных достигаются: пригодность/доступность данных; своевременность; точность, корректность и целостность; согласованность и ясность; завершенность; сокращение дублирования; ускорение обработки и сокращение времени ожидания; уменьшение стоимости сбора и хранения данных; всесторонняя обоснованность решения и интегрированное принятие решений [4].

Даная статья посвящена обоснованию необходимости использования репозитория как интегратора разнородных ресурсов геоинформационной системы (ГИС) и формулированию требований к архитектуре информационных ресурсов ГИС с участием репозитория.

План изложения материала следующий: в разд. 2 описывается роль централизованного репозитория в формировании геоинформационного пространства; в разд. 3 обсуждаются возможности репозитория в определении пространственных классов ограничения целостности геоданных, являющиеся важным аспектом управления качеством геоданных, а также архитектура и концептуальная модель репозитория, обеспечивающая контроль огра-

[1]Институт проблем информатики Российской академии наук, s.dulin@ccas.ru

[2]Научно-исследовательский и проектно-конструкторский институт информатизации, автоматизации и связи на железнодорожном транспорте (ОАО «НИИАС»), I.Rozenberg@gismps.ru

[3]ЗАО «ИнтехГеоТранс», umanvi@yandex.ru

ничений целостности со стороны широкого круга пользователей; в разд. 4 рассматривается композиционный подход к конструированию репозитория на основе формирования системы компонентных объектов ГИС, и завершается раздел обсуждением принятой в настоящее время архитектуры обработки геоданных с участием репозитория.

2 Репозиторий — интегратор данных о состоянии объектов и процессов

Пространственные данные составляют ядро большинства ГИС и оказывают наибольшее влияние на многие процессы принятия решений при управлении объектами. Данные объектов железнодорожного транспорта всегда идентифицируются своим географическим местоположениям и пространственными отношениями, поэтому ГИС и пространственный анализ данных являются важнейшими средствами поддержки процессов управления объектами.

В большинстве реализаций ГИС до настоящего времени пространственные данные сохранялись и обрабатывались в персональных или ведомственных базах геоданных, которые ограничивали совместное использование и редактирование данных. Возрастающие требования к совместной обработке пространственных данных для различных приложений выявили острую потребность в масштабируемости ГИС и создании геоинформационного пространства [1, 5].

В данной статье обсуждается возможность усиления роли ГИС в поддержке развития интегрированных систем управления объектами на базе централизованного репозитория. Под репозиторием здесь понимается предметно-ориентированная информационная корпоративная база данных, специально разработанная для поддержки принятия решений. Репозиторий строится на базе клиент-серверной архитектуры, реляционной системы управления базами данных (СУБД) и утилит поддержки принятия решений.

Ключевой задачей совершенствования средств обработки геоинформационного контента является создание геоинформационного пространства, позволяющего осуществить интеграцию пространственно-распределенной информации (семантической, метрической и топологической), с которой имеет дело ГИС, с данными высокоточного спутникового позиционирования, объединив их в единой геоинформационной базе данных отрасли. Это позволит представить все виды геоинформационных ресурсов отрасли в виде геореляционных структур, рассматривать их во взаимосвязи и оперативно получать информацию необходимого вида и содержания.

Создание геоинформационного пространства с концептуальной точки зрения может быть разделено на два основных направления:

(1) разработка принципов интеграции геопространственной информации на растровой и/или векторной основе с присоединенными базами фактографических данных, метаданных, данных высокоточного спутникового позиционирования, а также пространственно-определенной вербальной информации;

(2) разработка принципов и подходов к многоуровневому семантическому моделированию и согласованной интеграции геопространственной и пространственно-определенной вербальной информации с одновременным использованием электронных знаковых форм представления геотекстов, а также растровой и/или векторной формы их представления.

Геоинформационное пространство, интегрированное в прикладные геоинформационные и автоматизированные системы, способствует решению ряда задач, комплексно повышая уровень безопасности, бесперебойности и надежности функционирования железнодорожного транспорта.

Можно выделить три основных сегмента потребителей данных геоинформационного пространства:

(1) задачи проектирования, реконструкции и текущего содержания объектов;

(2) задачи управления процессами, базирующимися на различных технологиях, в том числе на технологиях спутниковых радионавигационных систем (СРНС);

(3) обеспечение систем безопасности функционирования координатной информации, которая служит вторичным информационно-управляющим контуром систем.

Каждый из этих сегментов предполагает разработку специальных методов интеграции геоинформационных ресурсов, обеспечивающих поддержку технологий управления железнодорожным транспортом. Более подробно с существующими железнодорожными технологиями можно ознакомиться в [6].

Актуализация геоинформационного пространства, поддерживаемого инструментальной средой ГИС, в современных условиях предполагает разработку и сопровождение централизованных репозиториев, основанных на интегрированных моделях

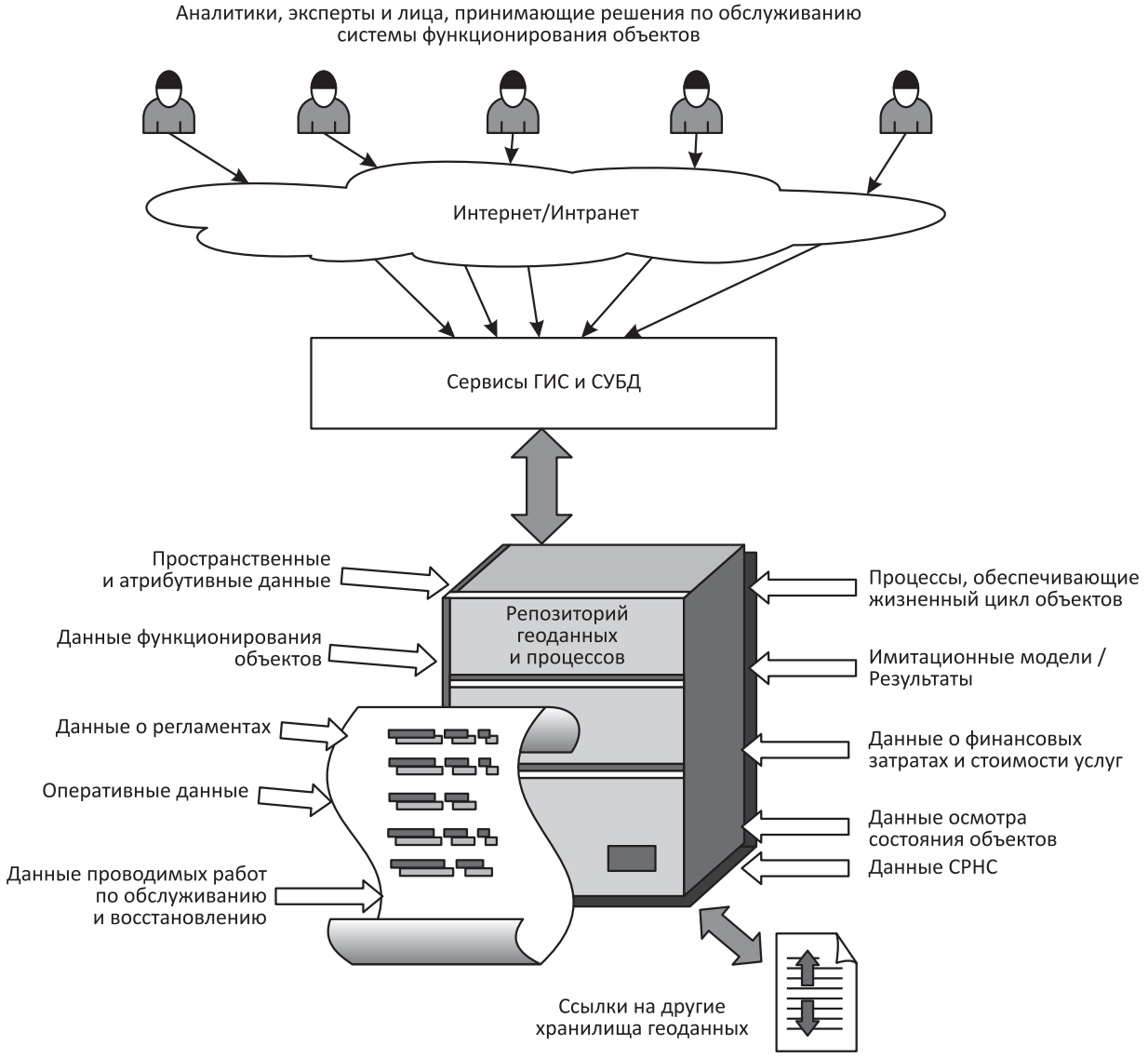


Рис. 1 Роль репозитория как интегратора данных о состоянии объектов и процессов

данных. Эти модели данных описывают характеристики объектов инфраструктуры и результаты функционирования подсистем объектов, отражая тем самым различные аспекты жизненного цикла объектов (рис. 1). Репозиторий может потенциально улучшить эффективность, рентабельность и координацию различных процессов управления объектами [7].

Централизованный репозиторий данных об объектах на основе ГИС построен как внешний модуль реляционной СУБД и поэтому может обеспечить широкий диапазон совместного использования данных, интеграции и сервисов управления, типа управления версиями, многопользовательского параллельного доступа и редактирования, безопасности и авторизации, а также сервисов метаданных. Репозиторий должен гарантировать согласованность и целостность данных и объединить различные форматы данных во всестороннее и непротиворечивое представление всей системы инфраструктуры. Обеспечивая единственный вход для доступа ко всем данным объектов, репозиторий может значительно улучшить сбор, организацию, управление и распределение данных всюду по жизненному циклу объекта. Интерфейс ГИС с репозиторием расширит возможности пользователей, позволив точно определить свои требования, сделать запрос и проанализировать данные объектов в пространственном контексте.

Репозиторий может обеспечить поддержку управления взаимозависимыми подсистемами инфраструктуры транспортных сетей интегрирован-

ным способом. Этими подсистемами инфраструктуры обычно управляют сепарабельно, при этом данные хранятся в отдельных и, возможно, несовместимых базах данных. Централизованный репозиторий может поддерживать перекрестные ссылки и отношения между различными подсистемами инфраструктуры, обеспечивающими взаимодействие между со-расположенными или накладывающимися объектами. Различные рабочие группы при этом будут в состоянии объединиться на основе согласованного представления данных.

Репозиторий допускает функциональную совместимость и эффективное совместное использование данных несовместимыми подсистемами управления объектами. Репозиторий может также поддержать использование распределенных источников данных и обеспечить доступ к этим источникам локально или через Интернет/Интранет (в доступной через сеть архитектуре клиент–сервер) распределенным приложениям-клиентам. Можно перечислить некоторые другие преимущества централизованных репозиториев:

- эффективное хранение, индексация, запрос и анализ данных объекта с параллельным поиском и редактированием этих данных для множественных приложений;
- поддержка координации и упрощение процессов управления объектами, увеличивающая операционную эффективность и расширяющая коммуникации между различными департаментами и заинтересованными пользователями;
- возможность многократного совместного использования данных и, таким образом, устранение дублирования усилий, потенциальной несогласованности и избыточности при сборе, верификации и хранении данных об объектах;
- использование согласованных, интегрированных и стандартизированных моделей данных и форматов;
- легко выполнимая интеграция инструментальных средств программного обеспечения, предварительно сформированных в виде отдельных модулей, в единую инструментальную среду.

3 Проектирование репозитория для управления метаданными ограничений целостности геоданных

Один из недостатков существующих коммерческих ГИС, предоставляющих настраиваемый набор услуг, — невозможность обеспечить адекватную операционную среду для конечных пользователей, которые являются экспертами в своей прикладной области, но обладают минимальным опытом в настройке программного обеспечения и формулировании требований к проекту базы геоданных [6]. Такие пользователи лишены возможности использования многих из особенностей коммерческой ГИС. Другой недостаток — отсутствие средств наложения ограничений целостности данных, что ставит под угрозу качество геоданных. Проектирование базы геоданных с контролем целостности при помощи существующих инструментальных средств требует знания некоторого языка сценариев. Конечные пользователи редко обладают этим типом знания.

Репозиторий ГИС предоставляет конечным пользователям средство для определения подмножества пространственных классов ограничения целостности геоданных без потребности в программировании [5, 7].

Если первые шаги в разработке базы геоданных для ГИС ставили в качестве главной цели приобретение данных и размещение их в релевантное место в системе, то в настоящее время акцент смещается в сторону эффективной организации и анализа геоданных, хотя никто не отрицает важности совершенствования технологии сбора геоданных. На ранней стадии разработки непространственных систем управления базами данных контролю целостности было уделено недостаточно внимания. В результате этого и пространственные наборы данных создаются с сомнительным качеством данных, что приводит к соответствующим результатам анализа, выполненного с использованием таких данных.

Учитывая высокую стоимость фиксации данных в формате ГИС, уже недостаточно неавтоматизированным способом обрабатывать определяемые пользователем ограничения целостности, которые влияют на качество данных.

Есть два основных подхода к проблеме качества данных. Первый — уменьшить ошибки в пространственных наборах данных, второй — вооружить пользователей знанием качества наборов данных и их содержания. Чтобы устанавливать ограничения целостности до ввода данных, необходимо включить ограничения целостности в язык определения данных базы данных так, чтобы они автоматически контролировались во время выполнения загрузки данных. Метаданные для этого процесса — словарь метаданных, описывающих характеристики, отношения и структуры. Каталог метаданных описывает происхождение и качество данных. Со-

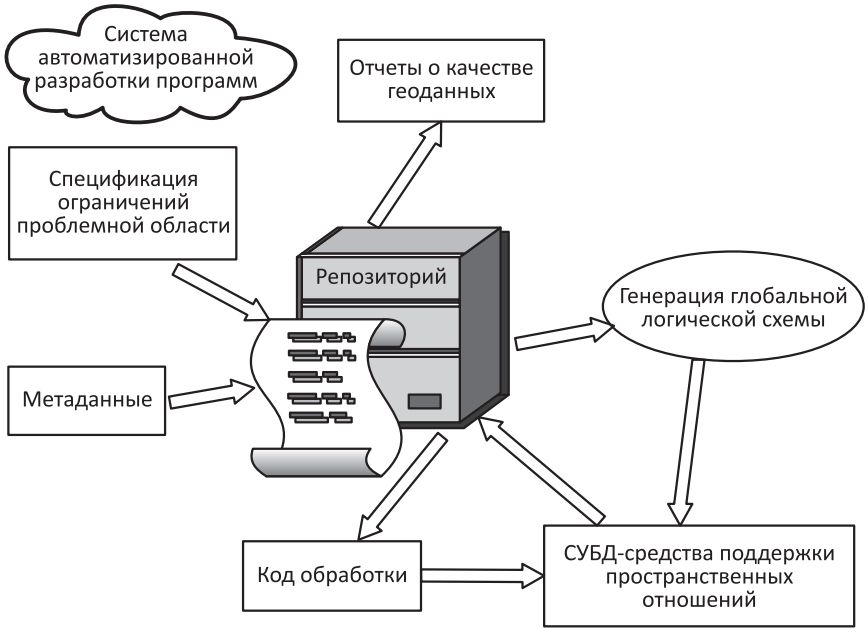


Рис. 2 Интегрированная пространственная среда программирования

поставление информации типа «кто ввел данные?» и «каково их происхождение?» представляет собой важную задачу при вводе данных для пользователя.

Репозиторий метаданных, который поддерживает хранение и обновление каталога метаданных, важен при проектировании геопространственных систем. Ограничения целостности могут быть неявно включены при проектировании, что позволит автоматически включить их в спроектированную систему. Пространственные и непространственные отношения также сохраняются вместе с ограничениями целостности и данными, касающимися преобразования объектов.

Репозиторий активизируется при вводе геоданных, контролируя соблюдение ограничений на данные или же фиксируя такие нарушения в файле регистрации нарушений.

Важно различать ГИС и пространственную информационную системную среду разработки. Дополнительная функциональность часто обеспечивается для разработчика приложения ГИС объединением ГИС с внешними программами или написанием специальной прикладной программы.

Особую важность в современных исследованиях представляет инструмент разработки ГИС, который позволяет пользователю специфицировать семантические ограничения на бинарные топологические отношения без потребности в программировании. Проектирование при этом не зависит от конкретной ГИС, но может быть связано с одной из них. Оно должно помочь пользователям в разработке пространственных приложений; один и тот же инструмент может использоваться, чтобы многократно строить проекты в различных масштабах. Одна из главных особенностей такого проектирования — создание репозитория для пространственных ограничений целостности, которые могут быть использованы во множестве приложений. Репозиторий отличается от различных словарей данных, поддерживающих ГИС, которые являются файлами или наборами файлов, ориентированных на одно конкретное приложение.

Интегрированная пространственная среда разработки программного обеспечения проиллюстрирована на рис. 2. Ключевой элемент в этой среде — репозиторий метаданных, который является средством контроля всех проектных изменений так же, как и репозиторий инструментальных средств разработки программного обеспечения в непространственных средах разработки.

При проектировании репозитория формулируются следующие цели:

1. Исследовать спецификацию правил и ограничений целостности в существующих пространственных и непространственных средах разработки.

2. Оценить существующую возможность инструментальных средств разработки ГИС представлять определяемые пользователем пространственные ограничения целостности.

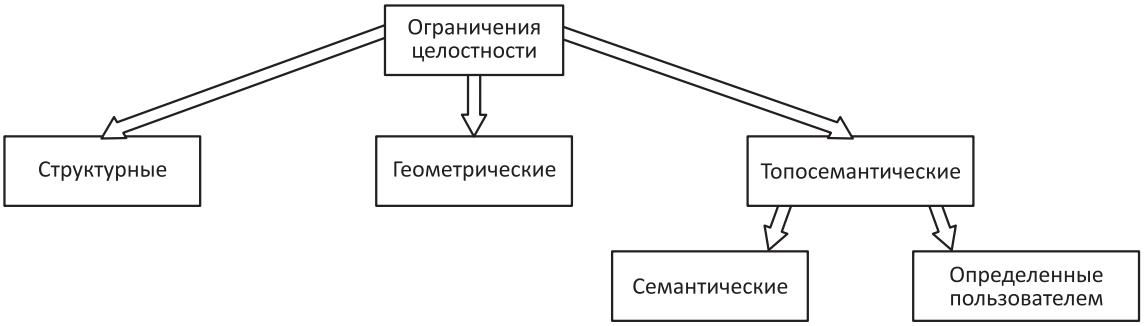


Рис. 3 Пространственные ограничения целостности

3. Определить методы регистрации и применения этих ограничений средствами, доступными для конечных пользователей.
4. Проверить эффективность этих методов.

Репозиторий хранит и контролирует подмножество классов ограничения целостности. Ограничения фиксируются двумя способами: посредством предложений языка определения данных СУБД и интеграцией с существующим программным обеспечением ГИС. Репозиторий хранит элементы модели геоданных, или данных о геоданных, средства генерации логических моделей данных из этих метаданных и поддержки проекта базы геоданных. В частности, должно быть средство для включения ограничений целостности в базу данных разрабатываемой системы. Метаданные качества и происхождения также сохранены в репозитории.

Проблеме качества пространственных данных, связанной с ограничениями целостности, в последнее время уделяется большое внимание. Ограничения целостности структурируются, как показано на рис. 3. Они подразделяются соответственно трем типам возможных ошибок: структурным, геометрическим и топо-семантическим.

Топо-семантические ограничения подразделяются на семантические ограничения и определяемые пользователем ограничения целостности.

Рисунок 4 иллюстрирует архитектуру предлагаемого репозитория. В качестве инструментальных средств можно использовать любой тип инструментальных средств, доступных в интегрированной среде разработки. Репозиторий управляет всем программным обеспечением в среде и представляет собой интерфейс между пользователем и инструментальными средствами.

Рисунок 5 представляет концептуальную модель репозитория, хранящего данные на метауровне. При проектировании неявно предполагается, что геометрия может иметь много графических представлений. Визуализация спецификаций включе-

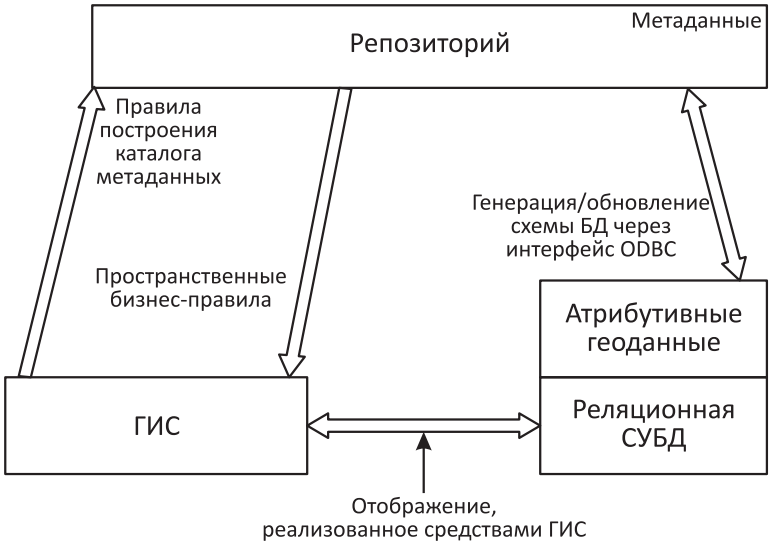


Рис. 4 Архитектура системы управления пространственными данными

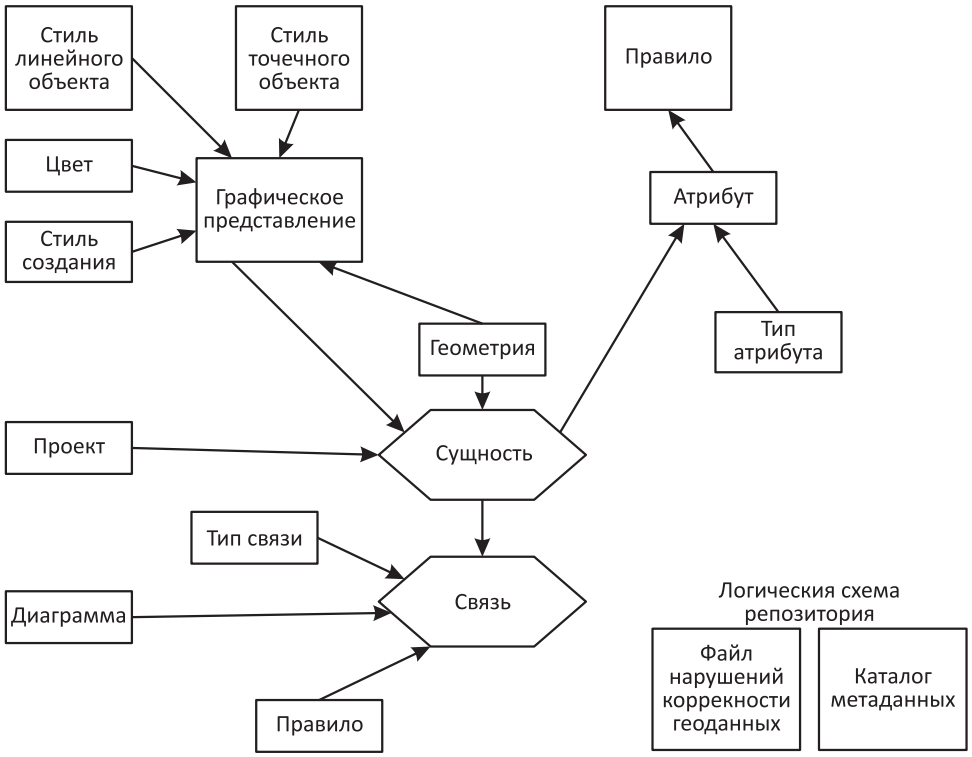


Рис. 5 Модель данных репозитория

на, чтобы позволить пользователю определить графические стили для объектов в ГИС, предполагая, что наличие визуальных команд вызова программы улучшило бы ясность относительно того, что собой представляют объекты, и таким образом уменьшило бы ошибку ввода данных. Это задача, которая непосредственно контролируется репозиторием. Для этого используются метаданные, касающиеся схематического изображения и использующиеся для генерации схемы, сохраняемой на том же самом метауровне.

Сущность диаграммы является внутренней для репозитория и содержит пространственную диаграмму связей. Логическая схема репозитория содержит таблицы хранения каталога метаданных, возвращаемых репозиторию от ГИС через программный интерфейс Open Database Connectivity (ODBC) доступа к данным, и файл нарушений корректности (целостности и согласованности). Правило (атрибута) иллюстрирует концептуальную модель хранения атрибутивных правил.

Центральная часть изображения модели репозитория содержит основную информацию для представления сущности и ее отношений. Репозиторий обрабатывает проекты через отношение сущность–проект этой части модели. Все записи в базе геоданных отмечены идентификатором проекта посредством значения внешнего ключа в сущности. В рабочем режиме репозиторий нужен для того, чтобы проекты совместно использовали допустимые правила. Но это требует дальнейшего развития модели, так как при этом необходима обработка отношений «многие к многим» между проектом и сущностью.

Топологические ограничения неявно поддерживаются средствами ГИС, с которой связан репозиторий. Правила представлены как предложения языка определения данных в языке структурированных запросов Structured Query Language (SQL) для правил атрибута и как структурированный текст для правил отношений (связи).

Система репозитория проектируется, чтобы облегчить разработку системы не вполне подготовленными пользователями. При этом главная задача — обеспечить спецификацию реальных объектов пользователями при существующих ограничениях на способ, которым данные об этих объектах могут быть введены. Эти ограничения задаются, чтобы управлять качеством данных. Система репозитория обеспечивает пользователям интерфейс, позволяющий устанавливать статические ограничения целостности на значения атрибута и определяемые пользователем ограничения целостности на пространственные отношения. Они автоматически

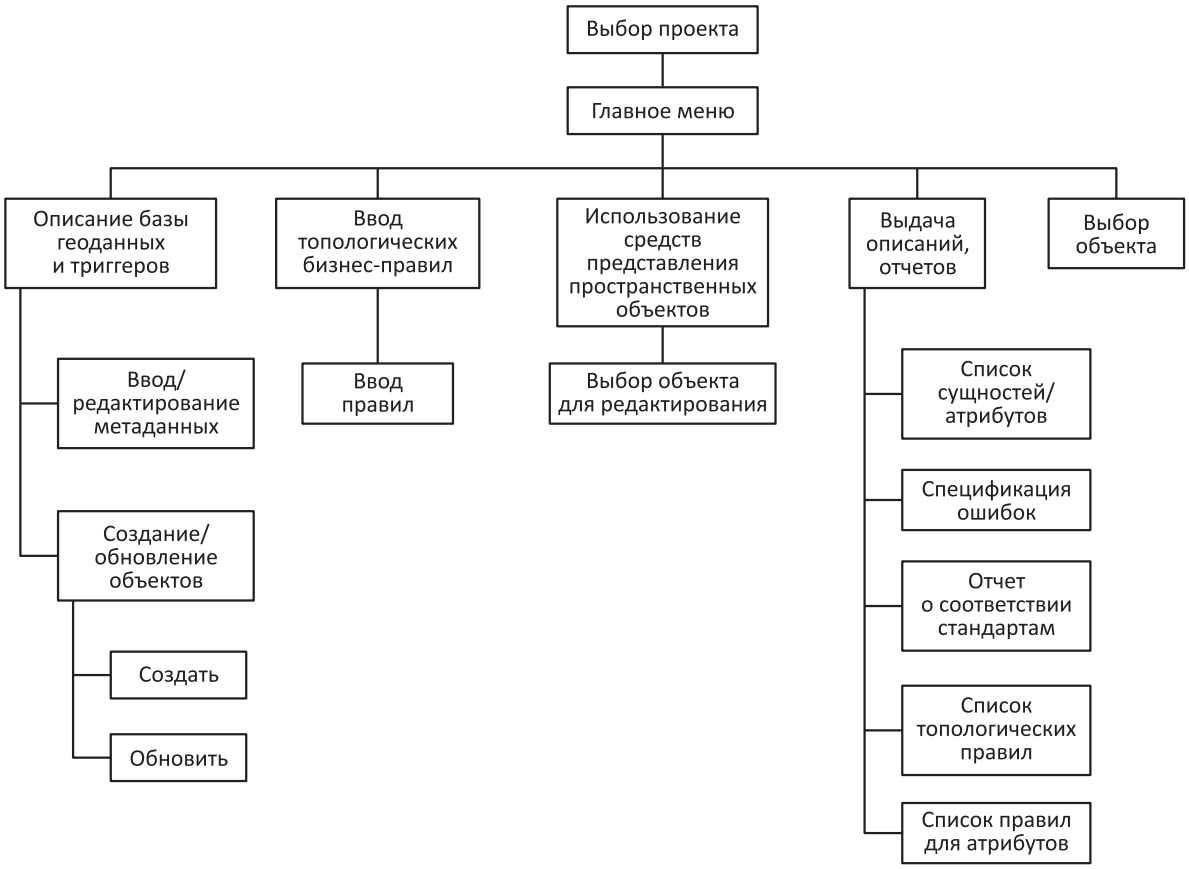


Рис. 6 Структура меню системы репозитория

перетранслируются в предложения языка определения данных или ограничения, выраженные как запросы к ГИС. Таким образом, не приходится нагружать программированием пользователей. Интерфейс репозитория также обеспечивает шлюз к ГИС. Управление в этом случае переходит к ГИС. Геоинформационная система способна реагировать на результат, который указывает, что имеет место нарушение правила. Такие нарушения при вводе геоданных приводят к откату входных данных или фиксируются в файле нарушений корректности репозитория. Кроме того, могут быть собраны автоматически и также предоставляться репозиторием метаданные об авторе, дате, пространственных границах, масштабе, проекции и системе координат.

На рис. 6 изображена структура меню системы репозитория. Цель операции по выбору проекта состоит в том, чтобы находить, вводить или изменять основные детали проекта. Эта операция начинается непосредственно после инициализации системы репозитория. Однако она может быть задействована на любой стадии выполнения проекта, когда пользователь пожелает перейти на другой проект. При этом вводятся детали проекта, включая координаты и масштаб.

Описание базы геоданных и триггеров (встроенных процедур) позволяет пользователю (1) описывать реальные объекты, их признаки и их графическое представление и (2) генерировать таблицы, представляющие эти объекты. Есть два обеспечивающих ее процесса: ввод/редактирование метаданных и создание/обновление объектов. В первом случае пользователь вводит имя сущности, ее атрибуты и правила, которые применяются к ним, ее геометрию и графический стиль, связанный с ней.

Функция ввода топологического бизнес-правила дает возможность пользователю определить пользовательские правила, ограничивающие отношения, в которых сущности могут принять участие [8]. Первоначально создается правило, основанное на сущностях, вовлеченных в связь и являющихся непосредственно связью.

Если заданы метаданные, которые сохраняются согласно существующим таблицам, то данные о правилах собираются на основе существующих в базе геоданных ограничениях целостности объек-

тов. Сообщение об ошибке формируется во время производимого пользователем выбора. Условия, на которых базируются ограничения целостности, сохранены в двух областях в репозитории. В первой условия сохраняются в соответствии с определенными атрибутами, а во второй — в соответствии с пространственными отношениями.

Когда пользователю требуются средства представления пространственных объектов, репозиторий инициирует запуск ГИС со всеми таблицами для выбранного открытого проекта. Репозиторий управляет всеми топологическими ограничениями и атрибутивными условиями. Ограничения проверяются ГИС при добавлении каждого пространственного объекта. Кроме того, при каждой транзакции проверяются все атрибутивные условия и топологические ограничения и в случае ошибок формируется соответствующее сообщение.

Существующие стандарты метаданных в пространственных информационных системах прежде всего отражены в каталоге метаданных. Существенная выгода использования подхода на основе репозитория в том, что репозиторий является активным и при разработке системы, и при ее эксплуатации.

Геоинформационная система обеспечивает сообщения (отчеты) о сущностях/атрибутах и дает список всех сущностей в проекте, их атрибутов и ограничений на эти атрибуты. Файл нарушений корректности геоданных дает список всех ошибок, которые произошли с пользовательскими правилами, идентификатор рассматриваемого объекта и его координаты.

Сообщения, обеспечиваемые репозиторием, включают также сообщения о топологическом правиле и о правиле атрибута. Топологическое сообщение о правиле включает все топологические правила проекта и условия выполнения этих правил. Сообщение о правиле атрибута включает атрибуты проектных сущностей, правила, связанные с ними, и текстовое правило, которое поставляется с сообщением об ошибке.

4 Конструирование системы репозитория компонентных объектов геоинформационной системы

Поскольку основной поток программной продукции в последнее время ориентирован на разработку отдельных модулей, удовлетворяющих требованиям модели компонентных объектов, программная индустрия в ГИС также становится модульно-ориентированной, обеспечивая как крупномасштабное развитие приложений ГИС, так и создание небольших гибких производственных систем. Для эффективной разработки таких систем необходимо глубокое исследование не только разработки компонентных объектов, но и управления ими.

Компонентные объекты ГИС перед каталогизацией в репозитории должны быть классифицированы на основе их специфики. Поэтому необходима система регистрации и поиска компонентов ГИС, в результате которой каталогизированный компонент позволит прикладным разработчикам найти нужный компонент или создать новый компонент посредством модификации и комбинации существующих компонентов и системы в целом. Основная цель разработки репозитория компонентных объектов ГИС состоит в том, чтобы обеспечить возможность использования во многих приложениях и функциональную совместимость разработанных компонентных объектов ГИС.

Чтобы реализовать ГИС более эффективно с точки зрения стоимости и времени, разработчики при конструировании репозитория компонентных объектов ГИС уделяют основное внимание возможности многоцелевого использования компонентных объектов и их функциональной совместимости. Это обусловлено тем, что большинство проектов ГИС представляет собой настройку на конкретное приложение типа административной информационной системы, интеллектуальной информационной системы, кадастровой информационной системы, системы управления в чрезвычайных ситуациях и адаптивной производственной системы. При этом каждое приложение нуждается в одних и тех же стандартных функциональных возможностях типа отображения геоданных и реализации запроса или специальных функциональных возможностях типа трехмерного средства просмотра и обработки данных системы GPS (Global Positioning System). Поэтому если существует универсальный репозиторий, хранящий компонентные объекты ГИС, и системные проектировщики или разработчики в режиме реального времени могут определить, где необходимый компонентный объект расположен, то они могут легко получить его, модифицировать или присоединить к своей системе.

Система компонентных объектов репозитория ГИС для регистрации и нахождения компонентных объектов ГИС строится в зависимости от их метаданных.

На рис. 7 показана общая концептуальная диаграмма системы компонентных объектов репози-

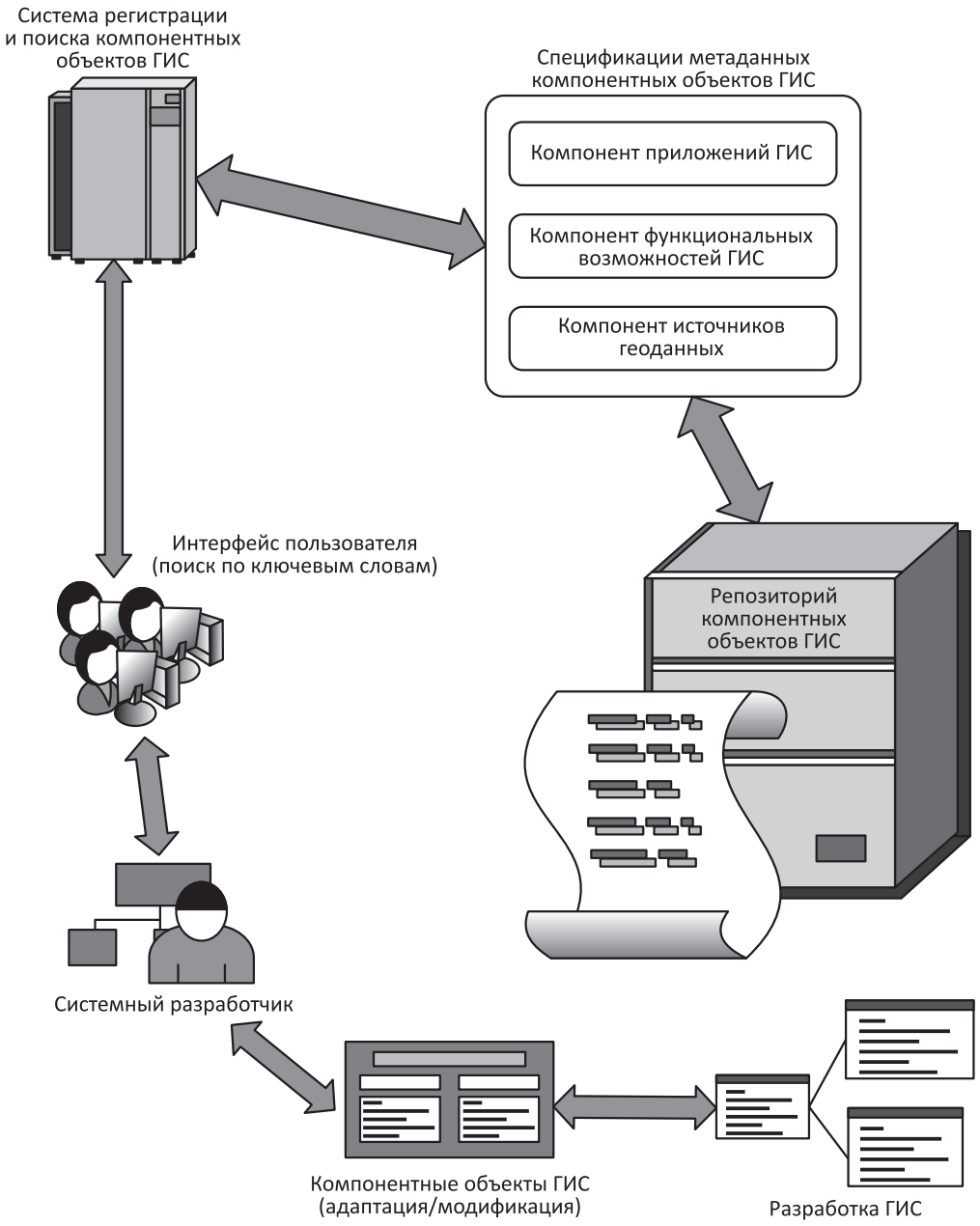


Рис. 7 Концепция исследования

тория ГИС. Пользователи вводят свои ключевые слова в интерфейсе поиска, и эти ключевые слова проверяются на соответствие логическому описанию СУБД через использование спецификаций метаданных компонентных объектов ГИС. Спецификации метаданных компонентных объектов ГИС подразделяются на три большие категории.

1. Первая категория — это компоненты источников геоданных ГИС, которые должны обеспечить функциональную совместимость пространственных форматов данных, находящихся в обращении в гетерогенной среде геоинформационного контента.

2. Ко второй категории относятся компоненты функциональных возможностей ГИС, которые могут использоваться как ядро ГИС при разработке определенного прикладного программного обеспечения ГИС. При этом следует заметить, что существуют как специфические функции, так и функции, общие для всех ГИС, такие как вывод на дисплей карты и атрибутов геоданных, или же анализ геометрических

характеристик, трехмерный анализ и авторизация.

3. Третья категория — компоненты приложений ГИС, которые указывают соответствие нескольким областям применения ГИС, типа административной информационной системы, системы управления ресурсами, системы управления в чрезвычайных ситуациях, интеллектуальной транспортной системы, городской информационной системы или кадастровой информационной системы.

Эта функциональная архитектура классификации позволяет указать место компонентных объектов ГИС и каталогизировать их в репозитории.

Рисунок 8 демонстрирует основную идею системного проекта с участием репозитория. На нем показана логика реализации системы в плане описания последовательности действий регистрации и поиска компонентных объектов ГИС.

В настоящее время данные о функционировании объектов передаются между различными программными инструментальными средствами, главным образом, двумя основными способами:

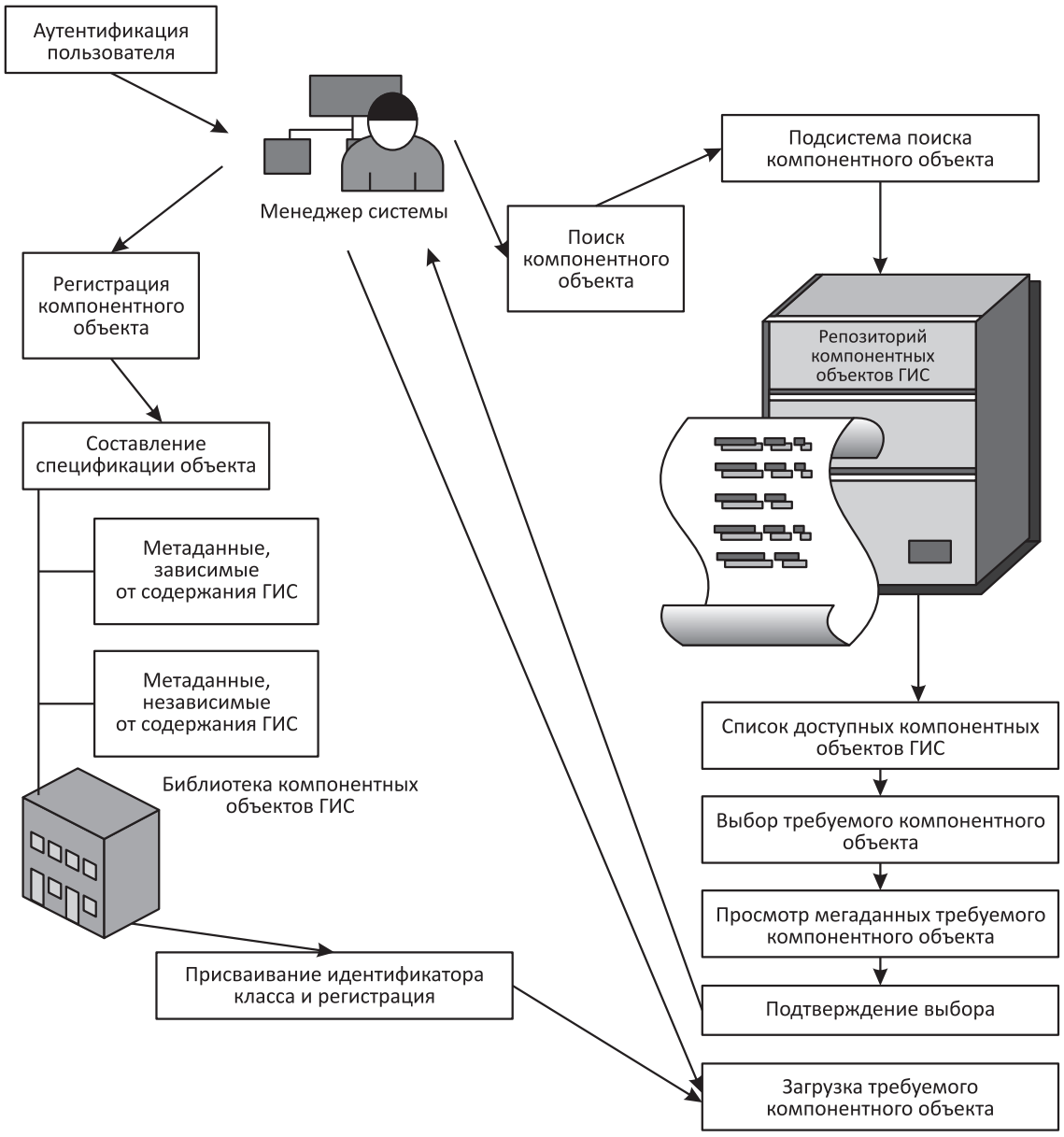


Рис. 8 Функциональная схема с участием репозитория

(1) с помощью трансляторов — чтобы преобразовывать данные при переходах между различными специализированными моделями данных и форматами. Процесс трансляции при этом связан с громадным объемом избыточного поиска данных, интерпретацией и повторным вводом данных и, как известно, допускает ошибки интерпретации и отображения результатов. Кроме того, специфика трансляторов налагает ненужные ограничения, требуя использования специализированных моделей данных и программных систем;

(2) с помощью ряда прикладных программных интерфейсов (API — applied program interface) — чтобы обратиться к внутренней модели данных приложения и вводить или извлекать данные непосредственно из приложения. Хотя некоторые из этих API фактически соответствуют промышленному стандарту, многие API являются специально разработанными и ориентированными под конкретную программную среду. Использование специально разработанных методов — очевидное препятствие на пути интеграции данных об объектах инфраструктуры.

Использование централизованных репозиториев, основанных на независимых от программной среды стандартных моделях данных, — по-видимому, самая жизнеспособная идея для интеграции данных управления объектами и программной функциональной совместимости.

Главная задача при формировании централизованного репозитория данных — разработать модель данных и соответствующую схему базы данных, чтобы представить данные жизненного цикла объектов объединенным, всесторонним и предпочтительно стандартизированным способом.

Программные инструментальные средства, соответствующие стандартам, могут быть легко интегрированы в репозиторий без необходимости разрабатывать специальные адаптеры для транслирования данных объектов в модель данных репозитория и, таким образом, будут способствовать развитию и развертыванию интегрированных систем управления объектами.

В отличие от традиционных моделей автоматизированного проектирования модели ГИС обеспечивают определение и использование семантически богатых объектно-ориентированных моделей, которые поддерживаются реляционной СУБД (РСУБД), предназначенной для хранения и управления атрибутивными данными. Объединяя пространственные и непространственные данные, модели ГИС допускают эффективную автоматизированную проверку корректности данных, гарантируя качество и надежность геоданных. Кроме того, архитектура клиент–сервер большинства ГИС дает возможность тонким клиентам эффективно обращаться к геоданным по сетям Интернет/Интранет, обеспечивая массовую публикацию пространственных данных в различных департаментах достаточно рентабельным способом.

Модель базы геоданных дает возможность выполнить проверку ограничений целостности на геоданные и использовать функции SQL реляционной СУБД для доступа к геоданным, обновления и управления транзакциями. Кроме того, модель позволяет определять специальные объекты, которые воплощают определяемую пользователем семантику, а также поддерживает сложные пространственные отношения типа сетей, топологий и ландшафтов.

Сервер приложений ГИС (рис. 9) представляет собой интерфейс, который позволяет управлять пространственными данными и хранить их. Важнейшее его преимущество состоит в возможности совместного доступа (чтения, записи, обновления, удаления) к используемым данным. Он распределяет пространственные данные для различного рода приложений, а также поставляет пространственные данные через глобальные сети по протоколу TCP/IP (Transmission Control Protocol / Internet Protocol).

Сервер приложений служит интерфейсом между ГИС и РСУБД для организации совместного доступа и управления пространственными данными как таблицами. В среде разнотипных баз данных, созданных различными организациями или отдельными пользователями, он обеспечивает общую модель хранения географической информации и значительно улучшает характеристики всей ГИС за счет распределения функций приложения ГИС между сервером базы данных и клиентом.

Сервер приложений управляет набором заданных таблиц (или системным словарем данных), которые хранят метаданные о пространственных данных, таких как пространственные ссылки, имена классов признаков и структур и пространственную индексацию.

Репозиторий поддерживает версионирование базы геоданных, что обеспечивает слежение за хронологией обновления геоданных и откат до прежнего уровня изменений, если в этом возникает потребность. Чтобы оптимизировать использование ресурсов памяти, изменения хранятся только в дельта-таблицах. Эти таблицы используются вместо копирования всей базы геоданных. Проведенные изменения в итоге приводят к одной версии, если все изменения согласованы.

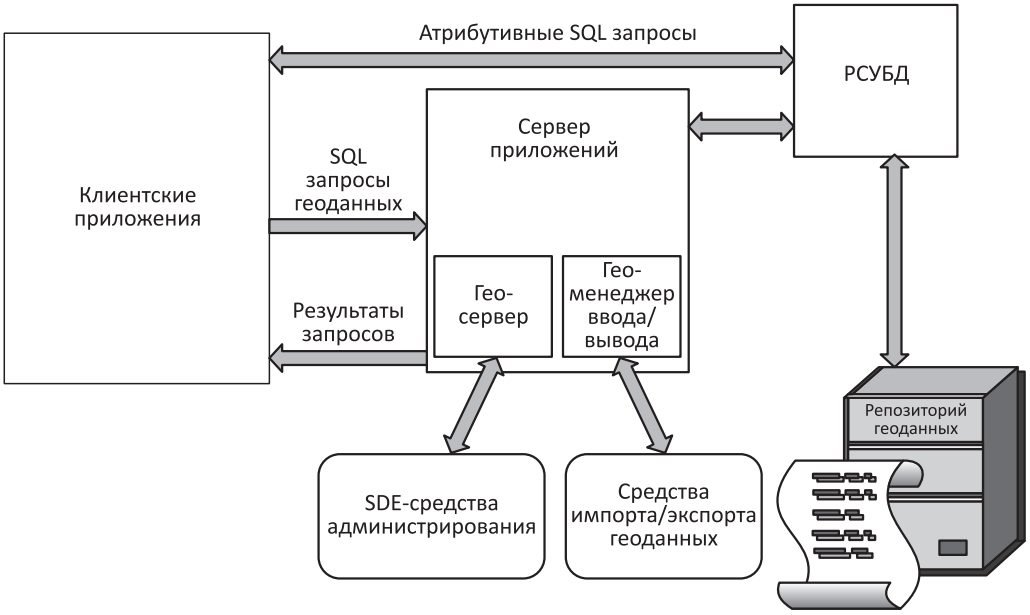


Рис. 9 Архитектура обработки геоданных с участием репозитория (SDE — Spatial Database Engine)

5 Заключение

В большинстве реализаций ГИС до настоящего времени пространственные данные сохранялись и обрабатывались в персональных или ведомственных базах геоданных, которые ограничивали совместное использование и редактирование данных. Возрастающие требования к совместной обработке пространственных данных для различных приложений выявили острую потребность в масштабируемости ГИС и создании геоинформационного пространства. В связи с этим следует обратить внимание на усиление роли ГИС в поддержке развития интегрированных систем управления объектами на базе централизованного репозитория.

В данной статье описан подход к проектированию репозитория для управления пространственными бизнес-правилами и его роль в разработке интегрированной среды программного обеспечения для ГИС. Следует отметить, что метаданные являются больше, чем средством каталогизации наборов данных. Они могут также эффективно использоваться при проектировании базы геоданных. Через контролируемый ввод геоданных репозиторий имеет возможность улучшить качество данных ГИС. Репозиторий активен при эксплуатации системы, проверяя ввод данных. Все нарушения ограничений регистрируются с выдачей сообщений о качестве геоданных. Репозиторий может обеспечить метаданные описания идентичности и происхождения наборов данных, введенных в систему. Эти два средства сообщения улучшают осведомленность о качестве рассматриваемого набора данных и поэтому предотвращают некорректное использование. Наконец, можно получить полный отчет о содержании репозитория, который помогает в администрировании базы данных. Качество геоданных в существующих ГИС часто невысокое. Существующие ГИС обеспечивают в лучшем случае только поддержку корректности ввода геоданных. Кроме того, инструментальные средства, используемые для разработки таких систем, не ориентированы на участие в разработке системы конечных пользователей. Основная цель представленной работы — определить главные особенности развития ГИС, допускающей конечных пользователей к участию в создании интегрированной среды, которая позволила бы пользователям задавать их собственные ограничения и получать качественные отчеты, соответствующие стандартам на метаданные.

Литература

1. *Розенберг И. Н., Дулин С. К.* Геоинформационный портал отрасли. Гарантировать достоверность данных // Железнодорожный транспорт, 2010. № 2. С. 12–17.
2. *Дулин С. К., Розенберг И. Н.* Об одном подходе к структурной согласованности геоданных // Мир транспорта, 2005. № 3. С. 16–29.
3. *Дулина Н. Г., Уманский В. И.* Структуризация проблемы улучшения пространственной согласованности баз геоданных. — М.: ВЦ РАН, 2009. 40 с.
4. *Дулин С. К., Розенберг И. Н.* Согласованное пополнение геоинформационного портала неструктурированны-

ми данными // Системы и средства информатики. — М.: Наука, 2005. Вып. 15. С. 194–218.

5. *Longley P. A., Goodchild M. F., Maguire D. J., Rhind D. W.* Geographic information systems and science. — 2nd ed. — New York: Wiley, 2005.
6. *Розенберг И. Н., Цветков В. Я., Матвеев С. И., Дулин С. К.* Интегрированная система управления железной дорогой / Под ред. В. И. Якунина. — 2-е изд., перераб. и доп. — М.: ИПЦ «Дизайн. Информация. Картография», 2008. 144 с.
7. *Baird M. P., Frome R. J.* Large-scale repository design // Cell Preservation Technol., 2005. Vol. 3. No. 4. P. 256–266.
8. *Orriens B., Yang J., Papazoglou M. P.* A framework for business rule driven service composition // 4th Workshop (International) TES 2003 / Eds. B. Benatallah, M.-C. Shan. — Springer, 2003. P. 14–27.

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2012. Т. 6. Вып. 2. С. 43–50

МЕТОДИКА МОДЕЛИРОВАНИЯ НАГРУЗКИ НА СЕРВЕР В ОТКРЫТЫХ СИСТЕМАХ ОБЛАЧНЫХ ВЫЧИСЛЕНИЙ

Д. В. Жевнерчук[1], А. В. Николаев[2]

Аннотация: Планирование серверного ресурса систем облачных вычислений является сложной задачей. Необходимо учитывать такие факторы, как состав и параметры аппаратной платформы, параметры системного программного обеспечения, управляющего выполнением прикладных программ, свойства трафика, порождаемого пользователями и определяющего режимы функционирования прикладных программ. Предложена методика оценки загрузки серверного ресурса открытых систем облачных вычислений (ОСОВ) на основе анализа процессов взаимодействия пользователей с программным обеспечением. Обоснована ее достоверность. Приведены результаты моделирования нагрузки на серверную часть системы управления средами имитационного моделирования.

Ключевые слова: облачные вычисления; имитационное моделирование; человеко-машинное взаимодействие

1 Введение

Вопросы проектирования систем поддержки удаленного вычислительного эксперимента до конца не решены. В ходе проектирования ОСОВ возникает задача моделирования потока запросов, приводящих к загрузке серверной части. Такие модели позволяют получить оценку аппаратного ресурса для обслуживания некоторого количества клиентских систем при рабочей и пиковой нагрузке. В работах [1–3] рассматриваются теоретические модели трафика в локальных и глобальных сетях. В основном полученные результаты имеют практическую значимость при проектировании средств передачи данных в сети. В работах [3–6] построены модели трафика, поступающего на вход серверов разного типа, таких как веб-серверы, серверы баз данных и др. На основании обзора работ были сделаны следующие выводы:

1. Модели описывают трафик систем, построенных на основе определенных технологий и/или предназначенных для решения ограниченного круга задач.
2. Процессы формирования запросов моделируются на основании замеров уже переданного в сеть трафика.
3. Модели описывают смешанный трафик.

Для анализа трафика ОСОВ классические методики моделирования трафика не эффективны, поскольку

- в общем случае ОСОВ обладает свойствами расширения по произвольным программно-аппаратным платформам, по решаемым задачам, по источникам нагрузки;
- в ОСОВ постоянно происходят качественные изменения, поэтому на основании конечного числа измерений трафика можно построить модели, описывающие ОСОВ только в некотором подмножестве состояний;
- отсутствуют развитые средства автоматизации и механизмы контроля перехода ОСОВ в новое качественное состояние.

Таким образом, задача разработки эффективных методик оценки нагрузки ОСОВ до конца не решена и является актуальной.

2 Постановка задачи

Была поставлена задача разработки методики моделирования и построения на ее основе моделей нагрузки на серверную часть в ОСОВ. К методике и моделям предъявлены следующие требования:

1. Модели должны описывать дифференцированный трафик, из которого можно выделить потоки, принадлежащие определенному программному обеспечению и связанные с определенными задачами.
2. Источник дифференцированного трафика должен определяться процессами человеко-машинного взаимодействия.

[1]Чайковский технологический институт (филиал) Ижевского государственного технического университета, drevnigeck@yandex.ru
[2]Чайковский технологический институт (филиал) Ижевского государственного технического университета, elodssa@yandex.ru

3. В модель должны передаваться эмпирические функции распределения вероятностей интервалов времени между передачами управляющих сигналов серверу, приводящих к существенной загрузке центрального процессора и оперативной памяти.
4. Должен проводиться системный анализ процессов решения пользовательских задач с применением программного обеспечения и учетом поведения пользователя при решении задач с помощью программ.
5. Должна обеспечиваться высокая степень автоматизации процессов сбора эмпирических данных.

Ставилась задача применить представленную методику для изучения системы обработки сред имитационного моделирования и проверить достоверность построенных моделей нагрузки на сервер.

3 Методика моделирования нагрузки на сервер

Предлагаемая методика моделирования нагрузки на сервер включает ряд этапов:

1. Сбор сведений о процессе взаимодействия клиента и сервера.
2. Определение последовательности выполнения действий.
3. Построение имитационной модели процесса взаимодействия.
4. Проведение экспериментов с имитационной моделью процесса взаимодействия и необходимой настройки.
5. Адаптацию полученного генератора нагрузки к работе с внешней системой.

Схема методики представлена на рис. 1.

На первом этапе необходимо собрать сведения о взаимодействии клиента и сервера. Клиент работает с сервером в режиме запрос–ответ. Необходимо получить данные об интервалах времени между определенными действиями клиента. Для упрощения сбора данных было разработано клиент-серверное приложение «Хронометр», позволяющее настроить список действий пользователя, требующих отметки времени выполнения. После настройки клиентская часть «Хронометра» начинает замерять интервалы времени, в течение которых выполняются действия, и передавать собираемые данные на сервер. Такой подход упрощает сбор

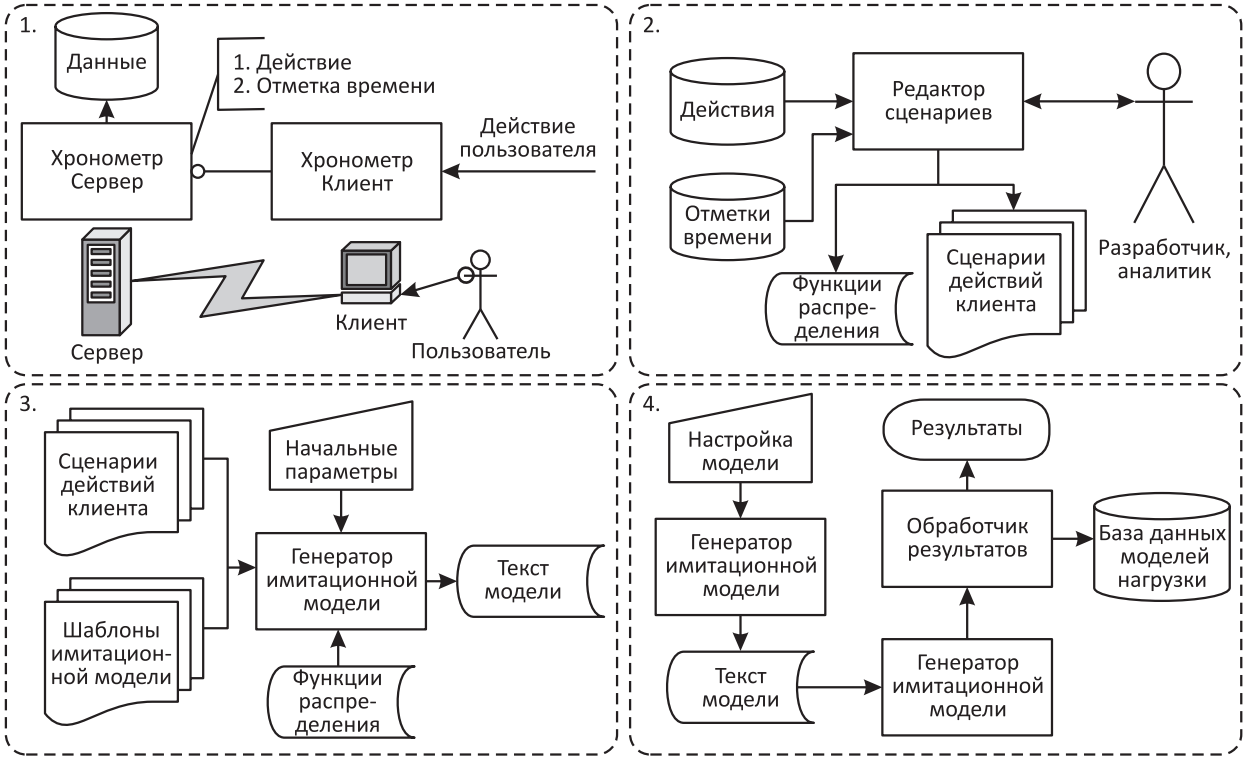


Рис. 1 Методика моделирования нагрузки на сервер

сведений, так как данные можно собирать сразу с группы пользователей.

На втором этапе определяется последовательность действий клиента по отношению к серверу. Для этого выполняются следующие шаги. На основе собранных данных строятся эмпирические функции распределения вероятностей интервалов времени, затраченного на действия клиента. Строятся сценарии человеко-машинного взаимодействия и вводится классификация пользователей, использующих определенные сценарии. Для упрощения работы было разработано программное средство «Редактор сценариев», позволяющее быстро обработать список действий пользователя и увидеть максимальное, минимальное и среднее время выполнения действий. Кроме того, с его помощью можно построить эмпирические функции распределения интервалов времени в синтаксисе языка GPSS (General Purpose Simulation System).

На третьем этапе строится имитационная модель процесса взаимодействия клиентов с сервером для оценки времени между событиями прихода запросов от клиента, приводящих к существенной загрузке центрального процессора и оперативной памяти. Для автоматизации построения имитационной модели было разработано программное средство «Генератор имитационной модели» (ГИМ). С его помощью можно быстро получить код модели на основании вводимых параметров, эмпирических функций и сценариев действий. Программное средство ГИМ использует заготовленные шаблоны на языке имитационного моделирования GPSS.

Далее проводится эксперимент с имитационной моделью, определение необходимых параметров и установка настроек. После этого строится модель трафика, поступающего на сервер, включающего запросы от клиента, приводящие к существенной загрузке центрального процессора и оперативной памяти. Все настройки и текст модели сохраняются в базу данных для последующего воспроизведения потока запросов.

На последнем этапе полученную модель генерации запросов пользователя адаптируют для работы с внешней системой.

4 Ход исследования

Исследования были проведены для среды моделирования GPSS World Student, с которой пользователи работают в режиме обучения. В построенных моделях трафика учитывалась информация о запросах, приводящих к прогонам имитационных моделей, что влияет на загрузку аппаратного ресурса. Была проверена гипотеза о достоверности построенных моделей.

4.1 Сбор данных о процессе взаимодействия учащегося со средой GPSS World Student

Был проведен хронометраж действий учащихся по изучению среды моделирования с помощью учебных моделей, по кодированию и отладке модели. Для формирования журнала действий пользователя использовалось программное средство «Хронометр 1.0».

Было замечено, что в ходе выполнения учебного задания учащийся сначала формирует код модели, далее модель тестируется и отлаживается. Это сопровождается определенным числом попыток компиляции модели и интервалами времени поиска и устранения ошибок в коде. После отладки выполняется разовый контрольный прогон модели, после чего анализируется выходной отчет, на основании которого осуществляется поиск логических ошибок. Учащимся может быть проведено несколько дополнительных исправлений кода. Модель вновь

Таблица 1 Наблюдение за процессом изучения GPSS World Student

Действие	Тип моделей	Категория	
		Успевающие	Неуспевающие
Количество ошибок компиляции в режиме отладки (1 задание)	Простые модели Сложные модели	[0–3] [4–12]	[2–6] [8–16]
Поиск ошибки и ее устранение, с		[30–120]	[90–200]
Анализ итогового отчета, с	Простые модели Сложные модели Первичное ознакомление	[20–60] [20–120] [60–120]	[40–120] [90–240] [90–120]
Кодирование новой модели, мин (подготовка первого варианта кода модели), мин		[120–300] [360–720]	[240–420] [600–1080]
Работа со средой моделирование по инструкции (время поиска функциональности)		[20–40]	[30–90]

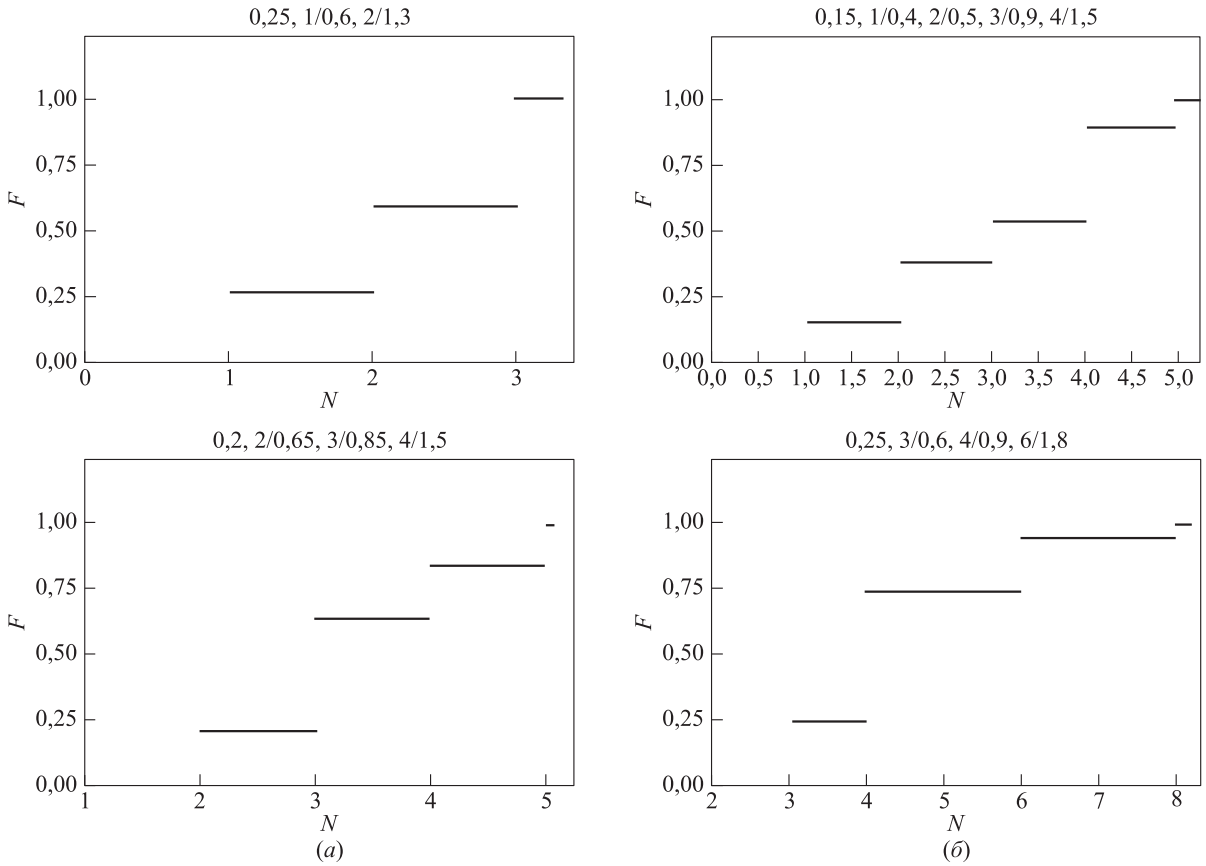


Рис. 2 Число шагов отладки простой (*а*) и сложной (*б*) модели успевающим (верхний ряд) и неуспевающим учащимся (нижний ряд)

тестируется, отлаживается, и результаты итогового прогона снова анализируются.

В режиме изучения готовой модели или среды моделирования с использованием методических указаний основное время тратится на анализ инструкций.

Наблюдения проводились за тремя учебными группами общей численностью 43 чел. Было проведено 5 занятий (10 академических часов). На основе полученных данных построена классификация учащихся и задач по времени решения. Все учащиеся разделены на две группы: «успевающие» и «неуспевающие», а задачи — на группы «простые» и «сложные». Полученные граничные оценки интервалов времени действий учащегося в режимах обучения и выполнения задания приведены в табл. 1.

С помощью программы «Редактор сценариев» были построены необходимые эмпирические законы распределения интервалов времени (рис. 2–5).

Полученные результаты были использованы при построении имитационных моделей взаимодействия учащихся со средой GPSS World.

4.2 Модель взаимодействия учащегося со средой GPSS World Student

Выделим события, приводящие к компиляции и прогону модели, а следовательно, и к загрузке центрального процессора. При возникновении события первого запуска модели (e_1) происходит компиляция, в результате которой формируется отчет о готовности прогона. При возникновении события «запуск модели» (e_2) выполняется прогон, в результате которого формируется отчет с откликом. В модели вводится 2 класса задач: простые и сложные, — и 2 класса учащихся: успевающие и неуспевающие. В зависимости от класса задачи и типа учащегося выбираются построенные в результате хронометража функции, определяющие количество ошибок компиляции и время задержки при написании кода модели, анализа ошибок компиляции, анализа итогового отчета. При построении модели были сделаны следующие допущения:

1. Время компиляции модели пренебрежимо мало по сравнению со временем прогона модели, а

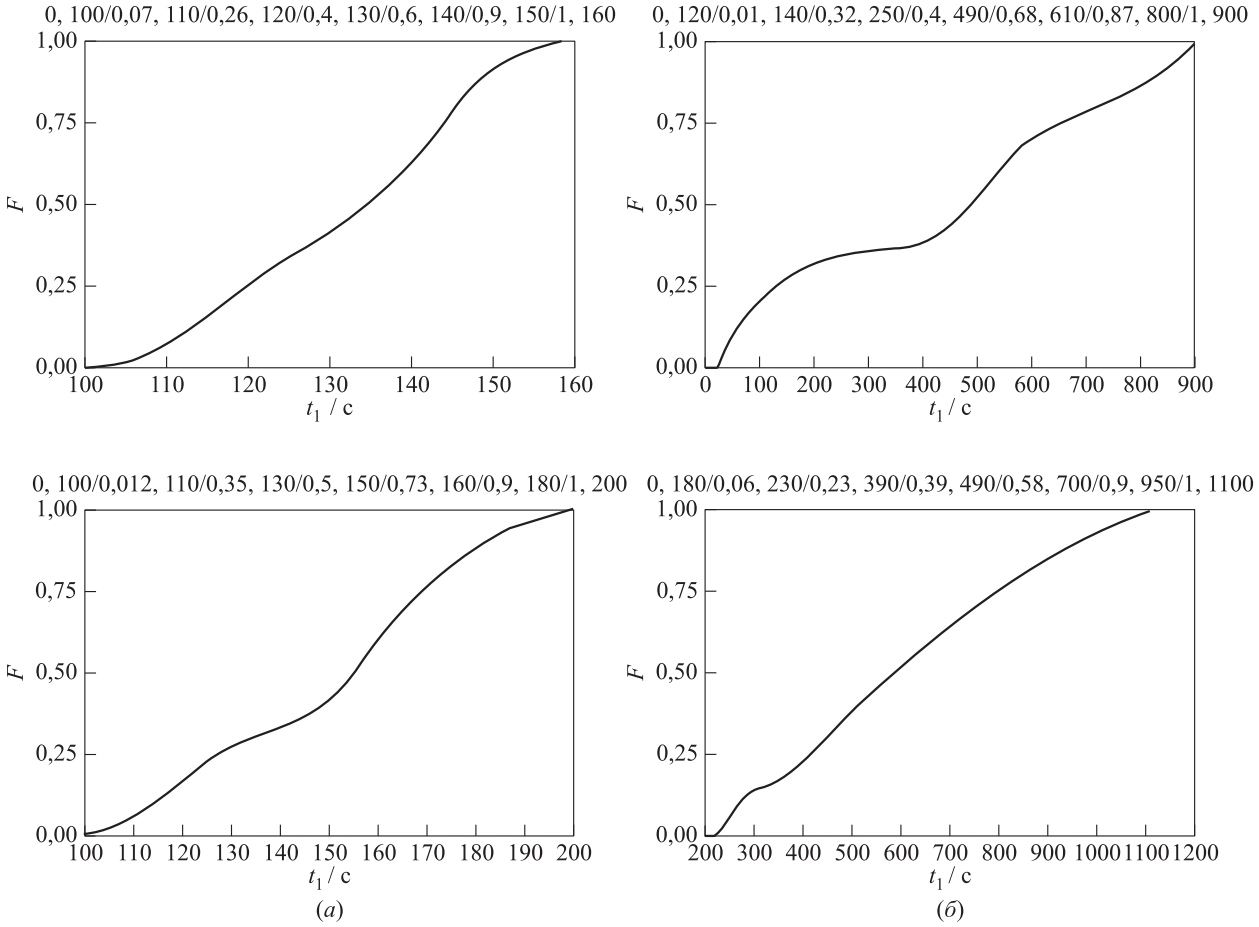


Рис. 3 Время написания первого варианта кода простой (*а*) и сложной (*б*) модели успевающим (верхний ряд) и неуспевающим учащимся (нижний ряд)

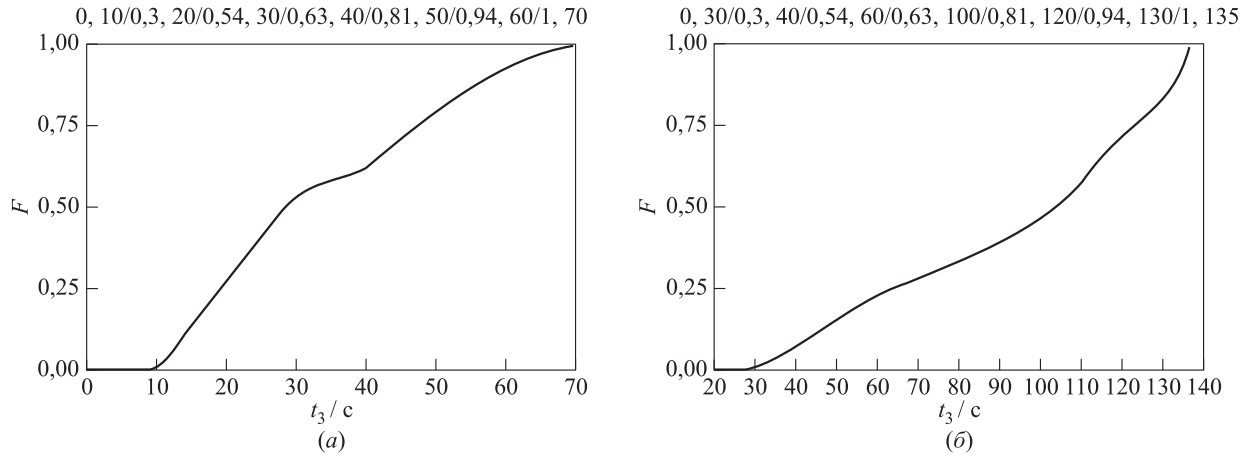


Рис. 4 Время поиска ошибок компиляции успевающим (*а*) и неуспевающим (*б*) учащимся

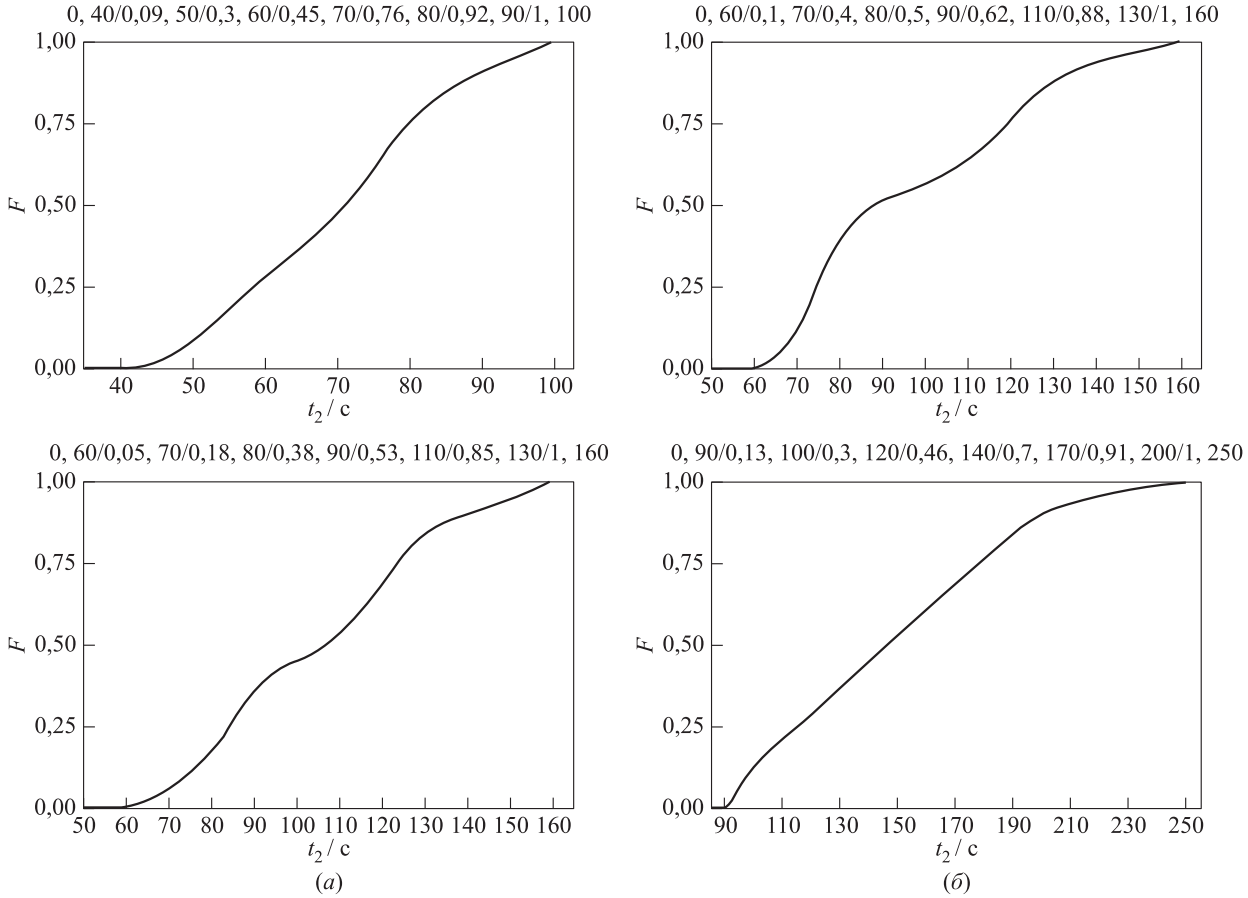


Рис. 5 Анализ отчетов простой (*а*) и сложной (*б*) модели успевающим (верхний ряд) и неуспевающим учащимся (нижний ряд)

также с периодами поиска, устранения ошибок и анализа отчета и составляет менее 0,1%.

2. Время прогона одной учебной модели варьируется в интервале [0,3–4] с в зависимости от алгоритмических свойств модели и от аппаратного ресурса, что также пренебрежимо мало в случае однопользовательского режима.

5 Сравнение отклика имитационной модели с реальной системой

Для проведения экспериментальных исследований была разработана система моделирования работы комплекса виртуальных лабораторий Open Virtual Research Space (OVRS), которая представляет собой клиент-серверное приложение для исследования процессов функционирования открытого виртуального исследовательского пространства (ОВИП) [7, 8].

Для оценки достоверности полученной модели был проведен хронометраж запросов на запуск имитационного эксперимента в среде GPSS тремя группами учащихся (табл. 2).

Таблица 2 Хронометраж запросов на запуск имитационного эксперимента

№ группы	Учащиеся		Задачи	
	Успевающие	Неуспевающие	Сложные	Простые
1	8	12	5	3
2	10	0	6	2
3	0	12	4	1

На рис. 6 приведены гистограммы пиковой загрузки реального и модельного сервера запросами на запуск имитационной модели.

Построены доверительные интервалы Велча для среднего значения числа заявок, поступающих в интервалы времени, равные 60 с: $(-0{,}21, 0{,}94)$, $(-0{,}83,\ 1{,}07)$; $(-1{,}17,\ 1{,}35)$.

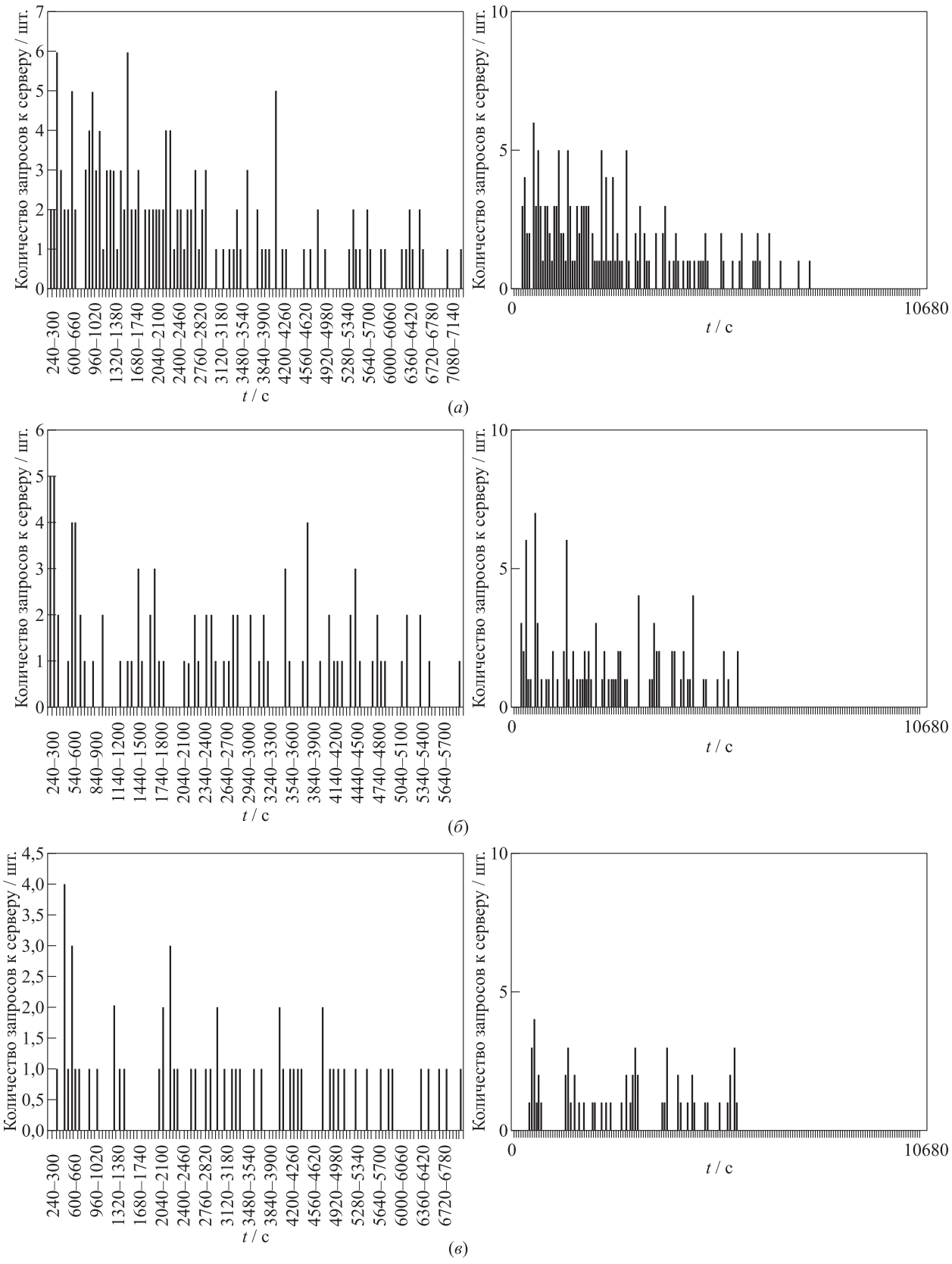


Рис. 6 Пиковая загрузка модельного (слева) и реального (справа) сервера по табл. 2: (*а*) для группы 1; (*б*) для группы 2; (*в*) для группы 3

Таким образом, достоверность построенной модели подтверждается наличием нуля в каждом интервале.

6 Заключение

В ходе исследования была разработана методика оценки нагрузки на серверную часть ОСОВ, особенностью которой является моделирование потока запросов на основе системного анализа поведения групп пользователей при решении определенных задач с применением программного обеспечения.

С использованием методики были построены имитационные модели и экспериментальная среда исследования взаимодействия пользователя с системой имитационного моделирования в режиме обучения.

Построены доверительные интервалы Велча для среднего значения числа запросов, поступающих на реальный сервер и его модель в интервалы времени, равные 60 с: $(-0{,}21, 0{,}94)$, $(-0{,}83, 1{,}07)$; $(-1{,}17, 1{,}35)$, что подтверждает достоверность построенной модели.

Использование методов системного анализа для исследования поведения групп пользователей при решении определенных задач с применением средств автоматизации позволяет адаптировать предложенную методику для изучения серверной нагрузки в ОСОВ.

Методика может быть применена при построении моделей процессов взаимодействия пользователей с произвольным программным обеспечением, доступ к которому предоставлен системами облачных вычислений. При этом модели будут отражать свойства потока запросов к серверной части, приводящих к загрузке ресурсов.

Перспективным направлением развития данного исследования является теория проектирования систем *cloud computing* и, в частности, вопросы зависимости пиковых загрузок от поведения пользователей.

Литература

1. *Willinger W., Taqqu M. S., Erramilli A.* A Bibliographical guide to self-similar traffic and performance modeling for modern high-speed networks // Stochastic networks: Theory and applications. — Oxford University Press, 1996. P. 282–296.
2. *Столлингс В.* Современные компьютерные сети. — СПб.: Питер, 2003. 782 с.
3. *Petroff V.* Self-similar network traffic: From chaos and fractals to forecasting and QoS // NEW2AN. — St. Petersburg, 2004. P. 110–118.
4. *Шелухин О. И., Тенякшев А. М., Осин А. В.* Фрактальные процессы в телекоммуникациях. — М.: Радиотехника, 2003. 480 с.
5. *Dang T. D., Sonkoly B., Molnar S.* Fractal analysis and modelling of VoIP traffic // 11th Telecommunications Network Strategy and Planning Symposium (International) (NETWORKS 2004) Proceedings. — Vienna, Austria, 2004. P. 123–130.
6. *Криштофович А. Ю.* Применение модели трафика сети ОКС № 7 для управления потоками сигнальной нагрузки // Инфокоммуникационные технологии, 2004. Т. 2. № 2. С. 25–27.
7. *Жевнерчук Д. В., Николаев А. В.* Открытый инструмент проведения дистанционного имитационного эксперимента // Вестник ИжГТУ. — Ижевск: ИжГТУ, 2008. № 2. С. 103–108.
8. *Ефимов И. Н., Жевнерчук Д. В., Николаев А. В.* Открытые виртуальные исследовательские пространства. Аналитический обзор. — Екатеринбург: Институт экономики УрО РАН, 2008. 83 с.

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2012. Т. 6. Вып. 2. С. 51–58

ЗАДАЧИ И ФУНКЦИИ БИБЛИОТЕК РАН В СОВРЕМЕННЫХ УСЛОВИЯХ

Н. Е. Калёнов[1]

Аннотация: Рассмотрены вопросы, связанные с изменениями в деятельности академических библиотек в связи с развитием сетевых технологий, бурным ростом числа электронных публикаций и баз данных (БД), доступных через Интернет. Показано, что и в современных условиях академические библиотеки являются неотъемлемой частью научной инфраструктуры. Традиционные задачи по информационному сопровождению научных исследований остаются прерогативой библиотек, однако должны выполняться на базе новых технологических решений с широким использованием современных сетевых технологий. Новые подходы к решению традиционных задач иллюстрируются на примере Библиотеки по естественным наукам Российской академии наук (БЕН РАН) (http://www.benran.ru). Наряду с традиционными задачами академические библиотеки готовы к решению новых задач, связанных с проведением библиометрических исследований, оцифровкой печатных изданий и т. п.

Ключевые слова: академические библиотеки; информатизация; автоматизация; обслуживание пользователей; информационное обеспечение; компьютерные технологии; автоматизированное рабочее место; электронные библиотеки; БЕН РАН

1 Введение

За последние годы произошли серьезные изменения в мировом информационном пространстве, которые непосредственно связаны с научными библиотеками. Бурное развитие электронных библиотек, широкое распространение новых издательских и библиотечных технологий, кажущаяся доступность через Интернет всей научной информации породили мнение, к сожалению, разделяемое рядом ведущих российских ученых, о бессмысленности существования академических библиотек. В этой связи представляется необходимым проанализировать их деятельность в современных условиях, рассмотреть перспективы их существования и развития.

С самого начала своего существования академические библиотеки решали две основные тесно взаимосвязанные задачи: (а) информационное обеспечение научных исследований и (б) сохранение знаний.

Первый вопрос, на который необходимо ответить: актуальны ли эти задачи в современных условиях?

Начнем со второй задачи, ответ по которой очевиден — знания необходимо сохранять. При этом в мире пока не придумано системы, альтернативной опубликованию результатов научных исследований, отражающих накапливаемые знания. Это могут быть публикации в научных журналах, монографиях, зарегистрированные патенты и авторские свидетельства. Под публикацией не обязательно рассматривать печатные материалы — это могут быть электронные носители, фото- или киноматериалы. Важно, что научный результат прошел экспертизу некоторым признанным научным сообществом — будь то редколлегия журнала, рецензенты издательств или ученый совет вуза.

Вплоть до настоящего времени сохранность знаний обеспечивают научные библиотеки, осуществляя комплектование своих фондов (в том числе и материалами на электронных носителях). Но если библиотеки перестанут существовать — кто и в каком виде будет сохранять знания? Теоретически это могут быть издательства или производители информации — ученые. Но издательства — коммерческие структуры, и их деятельность полностью определяется экономической конъюнктурой. Ежегодно в мире происходит слияние и разделение научных издательств, банкротство одних и появление других. Поскольку в сохранности знаний заинтересовано общество в целом, поручать эту серьезную задачу организациям, заинтересованным в получении максимальной прибыли, вряд ли целесообразно.

Еще более сомнительным представляется вариант обеспечения сохранности знаний их производителями — научными организациями или отдельными учеными. Очевидно, что при этом, во-первых, существенную роль будет играть субъ-

[1] Библиотека по естественным наукам Российской академии наук, nek@benran.ru

ективизм, а во-вторых, если ученые будут заниматься проблемами организации хранения и предоставления информации, им не останется времени на проведение собственно научных исследований и неминуемо встанет вопрос о создании для этого специальной структуры, которая фактически и будет являться библиотекой.

Возможен еще третий вариант, когда полученные знания формируются в электронном виде и загружаются в Интернет авторами, с тем чтобы любой желающий мог их оттуда «выудить». Этот вариант также не выдерживает серьезной критики, поскольку не обеспечивает экспертизу качества полученных результатов, оставляет открытым вопрос научного приоритета, требует наличия специальных служб, обеспечивающих поддержку и сохранность ресурсов (фактически аналогов библиотеки).

Таким образом, для обеспечения сохранения и предоставления знаний необходимо существование специальной структуры, которую логично отождествить с библиотечной системой, поскольку подавляющее большинство физических носителей знаний до настоящего времени хранится именно в библиотеках. Очевидно, что библиотеки как хранители знаний должны претерпеть серьезные изменения и вместе (а в дальнейшем, возможно, и вместо) с книжными стеллажами обладать мощными вычислительными средствами. Соответственно, и специалисты, работающие в библиотеках, должны получить необходимую квалификацию.

Перейдем к проблеме информационного обеспечения научных исследований. Нужны ли в РАН библиотеки для ее решения?

Оппоненты утверждают, что, поскольку вся важная научная информация имеется в Интернете, а научные издательства переходят на технологию *print on demand* (печать по требованию), надобность в библиотеках как структурах, обслуживающих информацией сотрудников РАН, отпадает — необходимую информацию каждый ученый может найти и получить самостоятельно через Интернет.

Предположим, что это так, но при этом возникает несколько проблем.

Первая из них — отбор и приобретение прав доступа к научным ресурсам (большинство ведущих мировых издательств и информационных центров предоставляют свои информационные ресурсы по подписке на платной основе). Очевидно, что отдельные ученые для себя лично этим заниматься не будут. Речь может идти о приобретении ресурсов для определенного коллектива исследователей уровня лаборатории, института, научного центра, отделения или РАН в целом. В условиях ограниченных финансовых ресурсов для оптимального решения этой проблемы необходима большая работа по анализу мирового информационного рынка, проведение переговоров с производителями (или поставщиками) ресурсов. Кроме того, как показывает практика, в процессе работы с удаленными ресурсами возникает множество технических и организационных вопросов, которые кто-то должен решать. Вряд ли ученым следует заниматься этими проблемами, которые отнимают достаточно много времени и требуют специальных навыков. Логично поручить эту работу библиотечным специалистам, которые во многих академических организациях ее уже выполняют.

Вторая проблема — работа с информационными ресурсами. До настоящего времени библиотечные работники РАН были «посредниками» между учеными и информационной средой — они осуществляли поиск ресурсов, отвечающих тематике исследований обслуживаемых групп ученых, информировали о появлении новых материалов и т. п. Если такие посредники исчезнут, то каждый научный сотрудник должен будет самостоятельно отслеживать появление новой информации, соответствующей тематике его исследований, а это, учитывая огромные объемы информационных потоков и их постоянный рост, потребует значительных затрат времени в ущерб собственно научным исследованиям.

Таким образом, роль академических библиотек в решении обеих вышеприведенных задач ничуть не уменьшилась по сравнению с предыдущими годами. Существенно должна измениться технология решения этих задач, что и будет показано ниже путем сравнительного анализа «традиционных» функций академических библиотек и их аналогов в современных условиях.

Традиционные функции академических библиотек включают:

- анализ информационных потребностей ученых;
- анализ мирового информационного рынка и приобретение необходимых научных ресурсов;
- формирование и доведение до пользователей (научных сотрудников) вторичной информации (сведений о поступивших изданиях, оглавлений журналов, аннотаций статей и т. д.);
- предоставление материалов (оригиналов изданий, копий статей) по запросам пользователей;
- организацию и хранение фондов.

Каждая функция подразумевает ряд основных процессов.

Рассмотрим традиционные и современные (применительно к практике БЕН РАН) способы их выполнения.

2 Анализ информационных потребностей ученых

Если исходить из того, что каждый отдельный ученый не в состоянии обеспечить себя самостоятельно в полной мере необходимой ему научной информацией и для этого требуется специальная структура, очевидно, что без знания потребностей обслуживаемых ученых эта структура действовать не может.

Согласно традиционной библиотечной технологи информационные потребности ученых, обслуживаемых данной библиотекой, отражаются в тематико-типологическом плане комплектования библиотеки (ТТПК), который представляет собой перечень тематических рубрик и типов изданий (научные, справочные и т. п.) по каждой из них. В XX в. ТТПК академических библиотек формировались в печатном виде библиотечными специалистами совместно с сотрудниками обслуживаемого научно-исследовательского учреждения (НИУ) и утверждались руководством НИУ. В дальнейшем библиотека отбирала литературу для своих фондов в соответствии с ТТПК. При изменении тематики исследований ТТПК перепечатывался. В централизованной библиотечной системе (ЦБС) БЕН РАН комплектование библиотек НИУ осуществляется централизованно, поэтому ТТПК отдельных библиотек передавались в центральную библиотеку (ЦБ), где на их основе формировался сводный печатный ТТПК ЦБС, который использовался комплектаторами при отборе литературы, подлежащей заказу. В современных условиях ТТПК остается попрежнему необходимым материалом для библиотек, отражающим информационные потребности ученых НИУ, однако ведется он в виде БД. В БЕН РАН разработаны специальные программные средства, обеспечивающие формирование ТТПК по отдельным библиотекам и сводного ТТПК по ЦБС в целом. Система обладает дружественным интерфейсом и полным формально-логическим контролем, обеспечивает унифицированный ввод информации в БД, предоставляет авторизованным комплектаторам развитые средства редактирования и выборки данных [1–3].

Для получения дополнительных сведений о реальных информационных потребностях пользователей библиотеки анализируют спрос на издания из своих фондов (в первую очередь спрос на журналы, поскольку полученные данные могут служить основой для корректировки подписки). Если раньше анализ спроса проводился вручную путем обработки читательских требований, то в современной системе БЕН РАН все заказы вводятся и обрабатываются в автоматизированном режиме, а для получения статистических данных и формирования оптимального заказа на журналы используются специально разработанные программные средства [4, 5].

Таким образом, принципиально процессы изучения информационных потребностей ученых остаются необходимой составляющей библиотечной деятельности, однако реализуются они на базе новых технологий, требующих соответствующей квалификации библиотечного персонала.

3 Анализ мирового информационного рынка и приобретение научных ресурсов

В XX в. основным источником информации об отечественном книжном рынке служили печатные тематические планы издательств (ТПИ), которые выпускались большими тиражами всеми издательствами страны и поступали в крупные библиотеки. Библиотеки ЦБС получали ТПИ через ЦБ, отмечали совместно с учеными необходимые издания и возвращали в ЦБ, где на их основе (после контроля по ТТПК) формировался и направлялся в издательства сводный заказ.

По зарубежным книгам БЕН выпускала ежемесячный указатель «Новые зарубежные книги», формируемый специалистами библиотеки на основе анализа материалов, поступающих из зарубежных издательств, и сопоставления их с ТТПК. Указатель рассылался и обрабатывался аналогично ТПИ. В масштабах страны информирование библиотек о новой зарубежной научно-технической литературе осуществлялось путем выпуска многотиражного издания, подготавливаемого ГПНТБ СССР на основе материалов, передаваемых крупнейшими библиотеками.

Сейчас такая технология представляется анахронизмом: ТПИ в печатном виде не издаются, а многие издательства присылают свои планы и прайс-листы в библиотеки по электронной почте. Кроме того, Российская книжная палата (РКП) формирует в электронном виде библиографическую информацию обо всех изданиях, поступающих по обязательному экземпляру. Используя эти данные, БЕН РАН разработала и внедрила в практику экспертную систему комплектования, основанную на сетевых технологиях [3, 6, 7]. Два раза в месяц в специальную БД, поддерживаемую на сервере БЕН РАН, вводится новая информация,

поступающая из РКП (на договорных условиях) и из ряда издательств (безвозмездно). Предварительно из общего массива данных программно отбираются издания, соответствующие сводному ТТПК, а также осуществляется сверка на дублетность (описания изданий, поступившие из РКП, могли ранее поступить из издательств). Авторизованные эксперты, официально выделенные руководством НИУ РАН из числа квалифицированных научных сотрудников, имеют возможность войти в БД, выбрать интересующую тематику и оценить издания с точки зрения целесообразности приобретения для библиотеки НИУ или ЦБС в целом. Полученные оценки обрабатываются специальными программными средствами и используются, в совокупности с ТТПК, для централизованного комплектования фондов библиотек НИУ.

Информация о зарубежном информационном рынке формируется специалистами БЕН РАН с использованием электронных каталогов издательств и базы данных Global Books-in-Print, конвертируется, загружается в экспертную систему и далее оценивается и обрабатывается аналогично отечественной информации.

Таким образом, пользователь участвует в отборе информации, подлежащей заказу; в определенном смысле реализуется технология, аналогичная традиционной, но на современном уровне. Очевидно, что эффективность такой технологии существенно выше традиционной, однако она требует от комплектаторов специальных навыков — они должны уметь работать с прикладными программными средствами, искать и обрабатывать информацию в Интернете, причем не только на русском, но и на других языках. Необходимо отметить, что работа по анализу мирового информационного рынка через сеть может быть существенно упрощена, если библиотека получит доступ к специальным БД, аккумулирующим издательскую информацию, однако доступ к подобным БД является платным и достаточно дорогим.

Что касается собственно приобретения ресурсов, то в современных условиях академические библиотеки все большее внимание уделяют приобретению прав доступа для своих пользователей к сетевым научным ресурсам. Поэтому наряду с традиционными процессами, связанными с получением материалов на физических носителях (заказ, контроль поступлений, регистрация и распределение изданий), библиотечные специалисты решают задачи, связанные с определением соотношения в приобретении печатных и сетевых ресурсов и организационным обеспечением доступа к последним (оформление лицензионных соглашений, сбор и передача поставщикам IP-адресов подключаемых НИУ, решение часто возникающих текущих проблем).

Как и другие направления деятельности, связанные с приобретением информационных материалов, процессы подключения пользователей к сетевым ресурсам определяются, в первую очередь, объемом финансовых ресурсов, выделяемых на эти цели. Права доступа к сетевым версиям подавляющего большинства серьезных научных журналов (так же, как и их печатные версии) предоставляются за плату, причем ее величина может существенно меняться в зависимости от того, с какого количества IP-адресов предоставляется доступ, сколько сотрудников имеется в данной организации и т. п. К сожалению, объемы финансирования, выделяемые на информационное сопровождение научных исследований в России (в частности, в РАН) и в большинстве стран, занимающихся наукой, несопоставимы. Так, годовой объем средств, получаемых БЕН РАН на информационное обеспечение около 120 институтов и научных центров, в 8 раз меньше средств, выделяемых на подобные цели библиотеке только одного факультета Гарвардского университета США. Как и любая коммерческая сфера, торговля доступом к научным ресурсам предусматривает различные скидки, нелинейную зависимость стоимости от количества пользователей и объема ресурсов и т. д. Проблемы с финансированием имеются практически во всех научных библиотеках мира, поэтому многие из них объединяются в консорциумы по доступу к информационным ресурсам, что позволяет экономить финансы для каждой из них. Достаточно подробно современные подходы к этим вопросам изложены в докладе американских библиотечных специалистов на последней конференции ИФЛА [8], где описываются подходы к экономии средств у двух консорциумов американских библиотек ORBIS и CIC, включающих соответственно 36 библиотек университетов и колледжей штатов Орегон и Вашингтон и 13 библиотек университетов северных штатов США. Объем финансирования этих библиотек колеблется в диапазоне от 6 до 20 млн долларов в год (цифры, несопоставимые с финансированием российских академических библиотек), тем не менее авторы утверждают, что средств на приобретение информационно-библиотечных ресурсов катастрофически не хватает. Задачу информационной поддержки своих пользователей они решают, сокращая до одного экземпляра подписку на печатные журналы внутри консорциума (или отказываясь вообще от печатной версии), координируя приобретение печатных книг, анализируя спрос на приобретенные ресурсы и корректируя подписку на основе этого анализа.

В этом направлении Россия нисколько не отстает от зарубежных стран. Подобный консорциум был организован в 1996 г. по инициативе БЕН РАН и включил РФФИ и 14 крупнейших библиотек страны. В рамках консорциума была организована скоординированная подписка на печатные версии журналов (финансируемая каждой библиотекой) и обеспечен доступ к научным журналам практически всех ведущих издательств мира (финансируемый РФФИ), которые передавали электронные версии журналов российской стороне, загружающей их на свои серверы и предоставляющей свободный доступ к ним всем желающим. С 2004 г. эта система перестала существовать по ряду причин (издательства отказались передавать информацию, а цены на онлайн-доступ существенно повысили, РФФИ изменил свою «политику», передав сопровождение уже загруженных массивов коммерческой структуре ООО «Научная электронная библиотека» — НЭБ). В настоящее время информация, поступившая в НЭБ до 2004 г., находится в свободном доступе на сайте www.elibrary.ru, все же новые журналы (в том числе и журналы РАН) предоставляются на коммерческой основе.

В последние годы РФФИ финансирует доступ своих грантодержателей к научным журналам основных зарубежных издательств, однако с 2012 г. эта деятельность также в определенной степени сворачивается (в частности, РФФИ из-за нехватки финансовых средств отказался от приобретения доступа к журналам второго по значимости в мире издательства Springer-Verlag).

В этой связи перспективы развития системы доступа к научным ресурсам со всех рабочих мест сотрудников РАН представляются достаточно туманными. Централизация же библиотек (когда структурные подразделения центральных библиотек располагаются в институтах и научных центрах и обслуживают их сотрудников) позволяет существенно экономить средства на приобретение доступа к журналам и БД путем ограничения числа рабочих мест компьютерами библиотеки. Именно такой подход в сложившихся условиях применяет БЕН РАН, обеспечивая в рамках выделенных ассигнований доступ к максимальному количеству важнейших журналов с компьютеров ЦБ и библиотек — ее отделений в НИУ РАН.

Острота проблемы доступа к важнейшим научным журналам прошлых лет, вероятно, в ближайшие годы будет снижена благодаря тому, что Минобрнауки планирует профинансировать приобретение их электронных архивов, размещение на отечественных серверах и доступ к ним всех граждан России.

4 Формирование и доведение до ученых вторичной информации

На протяжении многих десятилетий академическими библиотеками отрабатывались наиболее эффективные формы и методы информационного обслуживания ученых. Обеспечивающие его процессы можно разделить на две группы — информирование об изданиях, имеющихся в библиотеках (раскрытие фондов), и поиск вторичной информации по тематике исследований ученых. Первая группа традиционно включала в себя карточные (реже — печатные) каталоги фондов (алфавитные, систематические, предметные), печатные указатели подписки и бюллетени новых поступлений. Вторая группа включала выпуск текущих и ретроспективных указателей литературы по различным тематическим направлениям, ведение тематических картотек, избирательное распространение информации (ИРИ), поиск библиографии по заданной пользователями теме и т. п.

В современных условиях на смену карточным каталогам пришли электронные, обеспечивающие все виды библиографического поиска; вместо печатных указателей на сайтах академических библиотек представлены разделы, отражающие новые поступления изданий. Наряду с информацией об изданиях, имеющихся на физических носителях, на сайтах академических библиотек представлены сведения о сетевых ресурсах, доступных их пользователям. Так, на сайте БЕН РАН (http://www.benran.ru) выделен раздел «Электронные версии книг и сериальных изданий, доступные пользователям БЕН РАН», который содержит подразделы «Сериальные издания», «Книги», «Справочники и энциклопедии»; в сводном каталоге журналов содержится перечень всех выпусков каждого журнала, поступивших в ЦБС БЕН РАН с 1990 г., а также ссылки на сайты журналов, перейдя по которым пользователь может знакомиться с оглавлениями журналов, аннотациями статей и читать полные тексты (если имеет соответствующие права). В каталоге представлены не только журналы, выписанные непосредственно БЕН, но и журналы, доступ к которым предоставлен ученым в рамках электронной библиотеки РФФИ и через Национальный электронный информационный консорциум (НЭИКОН), реализующий соответствующую программу Минобрнауки.

Что касается текущего и ретроспективного информационного обслуживания, то, как показывают опросы пользователей, эти процессы по-прежнему

востребованы учеными, но формы их реализации существенно изменились — вместо тематических картотек и указателей создаются проблемно-ориентированные БД; информацию для удовлетворения запросов пользователей библиотечные специалисты ищут в Интернете и результаты поиска направляют пользователям по электронной почте. Многие библиотеки создают на своих сайтах разделы, содержащие ссылки на сетевые ресурсы по тематике исследований своих институтов. В качестве примеров наиболее активных в этом направлении библиотек ЦБС БЕН РАН можно привести Центральную библиотеку Пущинского научного центра (ПНЦ) РАН (http://cbp.iteb.psn.ru) и библиотеку Математического института им. В. А. Стеклова РАН (http://libserv.mi.ras.ru).

БЕН РАН поддерживает на своем сайте раздел «Естественные науки в Интернете», который представляет собой набор «метауказателей» ресурсов по основным разделам естественных наук. Каждый метауказатель (например, «Физика в Интернете», «Биология в Интернете» и т. п.) содержит сведения об указателях ресурсов, представленных в Интернете по соответствующим разделам науки — дается описание указателя, сведения об организации, формирующей указатель, и ссылка на него. Специальные сотрудники библиотеки обеспечивают поддержку метауказателей в актуальном состоянии.

В современных условиях одной из информационных функций сотрудников академических библиотек (по аналогии с традиционными указателями новых поступлений) должно стать отслеживание новых выпусков журналов, появляющихся в Интернете, к которым имеется доступ у сотрудников обслуживаемых НИУ. Библиотечные работники должны регулярно просматривать сайты журналов, соответствующих информационным интересам ученых, и информировать их (по электронной почте или на сайте) о выходе нового выпуска. Такой сервис, ориентированный на подразделения НИУ, позволяет существенно экономить время научных сотрудников, высвобождая его для исследовательской работы.

Одна из традиционных функций академических библиотек — ведение картотек трудов сотрудников. В современных условиях картотеки заменяются БД. На сайте БЕН РАН представлены БД публикаций сотрудников ряда НИУ РАН, поддерживаемые библиотечными специалистами с помощью унифицированного программного обеспечения, разработанного в БЕН. Потребность в таких БД не только не уменьшается, а существенно возрастает в связи с введением критериев оценки эффективности научной деятельности, в том числе связанных с количеством и качеством публикаций. В этой связи необходимо отметить роль библиотек в проведении библиометрического анализа публикаций сотрудников институтов с использованием мировых БД, таких как WEB of Science и SCOPUS. Многолетний опыт работы с подобными БД, имеющийся в БЕН РАН и других центральных академических библиотеках [9–12], показывает, что для достижения корректных результатов как по отдельным авторам, так и по коллективам ученых необходимо проведение достаточно серьезной и большой работы, требующей знания специфики данной БД. Дилетантский подход при оценке публикационной активности приводит к ошибочным выводам, которые могут быть чреваты серьезными последствиями для науки. Библиометрический анализ публикаций ученых РАН должны проводить квалифицированные специалисты, для которых эта деятельность является профессиональной. В РАН такими специалистами являются, в первую очередь, сотрудники библиотек. Наиболее активную работу в этом направлении в ЦБС БЕН РАН ведет Центральная библиотека ПНЦ, о которой уже шла речь выше. Ее сотрудники периодически проводят анализ публикаций ученых НИУ ПНЦ и представляют на своем сайте результаты этого анализа.

5 Предоставление материалов по запросам пользователей

Традиционные формы обслуживания читателей академическими библиотеками постепенно уступают место новым технологиям. В БЕН РАН все заказы на оригиналы изданий и копии материалов принимаются только в автоматизированном режиме. Каждый авторизованный читатель может сформировать свой заказ через Интернет или через компьютер в зале каталогов ЦБ. Сотрудники библиотеки регистрируют в базе данных все отказы, поступающие из отдела фондов. Это позволяет с помощью специальных программных средств проводить 100%-ный анализ спроса на литературу и причин отказов, что, в свою очередь, является основой для корректировки комплектования ЦБС и перераспределения изданий между библиотеками внутри единого фонда. Если 10 лет назад основными видами выполнения заказов являлось предоставление оригинала издания или ксерокопии статьи, то сейчас значительная доля заказов выполняется в виде электронных копий. Действующая в БЕН РАН система электронной доставки документов позволяет авторизованному пользователю осуществлять заказ непосредственно из Интернет-каталогов, представленных на сайте библиотеки,

и получить материалы в течение нескольких часов.

Наряду с предоставлением материалов из собственных фондов академические библиотеки (как центральные, так и многие библиотеки НИУ) организуют доступ своих пользователей к сетевым информационным ресурсам. Как показывает практика (подтвержденная результатами анкетирования, проведенного БЕН РАН среди своих пользователей), многие пользователи – сотрудники РАН испытывают определенные затруднения при работе с сетевыми ресурсами (что усугубляется разными интерфейсами, принятыми у разных поставщиков). Поэтому одной из функций академических библиотек должно стать обучение пользователей работе с сетевыми ресурсами. Эта функция, как и многие другие, не является чем-то новым для библиотек — раньше они обучали читателей пользованию каталогами, теперь — пользованию сетевыми ресурсами. Меняются средства работы библиотечных специалистов, повышаются требования к их квалификации, но их функции как «проводников» ученых в мировом информационном пространстве остаются.

6 Организация и хранение фондов

Пока продолжается выпуск научных материалов в печатной форме, традиционные функции библиотек, связанные с обеспечением сохранности литературы, созданием условий для быстрого поиска и выдачи нужного издания сохраняются. В современных условиях к ним добавляются процессы, связанные с обеспечением сохранности электронных носителей информации. Здесь достаточно много неясностей — нет четкого представления о том, как часто надо перезаписывать данные, чтобы защититься от их физического износа; как обеспечить возможность считывания данных с таких носителей, как CD- или DVD-диски, через несколько лет после записи, когда принципиально поменяются носители и технология записи на них данных (например, уже сейчас достаточно трудно найти компьютеры для чтения 3,5-дюймовых дискет, не говоря уже про 5-дюймовые); как долго надо хранить электронные издания (печатные издания в большинстве библиотек периодически списываются) и т. д. Технически и программно многие из этих вопросов решаются, но необходимо отдавать отчет в том, что если академические библиотеки будут оцифровывать и хранить электронные издания, то они должны обладать соответствующими техническими ресурсами.

Проблемы, связанные с электронными научными ресурсами, должны решаться на серьезном уровне, а не декларациями о переводе всех фондов библиотек в электронную форму в течение ближайших лет. К этим проблемам вплотную примыкают вопросы, связанные с формированием электронных библиотек (ЭБ) путем оцифровки печатных изданий. Многие традиционные библиотеки мира участвуют в формировании ЭБ, реализуются международные проекты в этой области. К наиболее успешным можно отнести проект Европейской электронной (цифровой) библиотеки Europeana, реализованный при поддержке Европейской комиссии в конце 2008 г. (http://www.europeana.eu/portal). В этой ЭБ представлены ресурсы, отражающие культурное наследие различных стран, в том числе (хотя и в небольшой степени) России.

Что касается академических библиотек России, то они активно вовлечены в работы по созданию ЭБ «Научное наследие России» (http://e-heritage.ru), которые ведутся в рамках одноименной целевой программы Президиума РАН с конца 2006 г. [13, 14]. Библиотеки формируют и вводят в ЭБ разнородную информацию о российских ученых, оцифровывают наиболее важные их работы, опубликованные в виде монографий и статей. Наполнение ЭБ осуществляется в хронологическом порядке начиная с XVIII в., что в первую очередь связано с законодательством об охране авторских прав. Согласно 4-й части Гражданского кодекса РФ библиотека не имеет права оцифровывать печатные издания и предоставлять их в сетевой доступ без письменного согласия авторов и других владельцев авторских прав, если со дня смерти автора прошло менее 70 лет. Тем не менее создание электронных библиотек является одним из перспективных направлений деятельности академических библиотек.

7 Заключение

Выше были перечислены функции, которые должны выполнять академические библиотеки в современных условиях, оставаясь востребованными учеными. Многие из этих функций в той или иной мере развивают традиционные информационно-библиотечные технологии. Но появился ряд новых задач, которые с успехом могли бы выполнять библиотеки. Это уже упомянутые работы по проведению библиометрических исследований, работы по созданию электронных библиотек. Кроме того, одним из направлений деятельности академических библиотек могла бы стать поддержка в актуальном состоянии информации, представленной

в Едином научном информационном пространстве (ЕНИП) РАН. Разработанная программная оболочка ЕНИП «Научный институт РАН» [15] позволяет вводить, редактировать и осуществлять многоаспектный поиск данных о деятельности НИУ РАН, в том числе о публикациях сотрудников с возможностью подключения полных тексов. Работа эта требует значительных усилий и временны́х затрат. По мнению автора, ЕНИП РАН не получает должного развития в том числе и из-за отсутствия в НИУ РАН сотрудников, обеспечивающих ввод и поддержку в актуальном состоянии всего информационного комплекса. Эта работа «не по профилю» ни системному администратору, ни администратору сайта института. В ряде НИУ она возложена на ученых секретарей, но у многих из них есть свои научные интересы, и поддержка информационных ресурсов является для них второстепенной задачей. Кроме того, объем работы в этом направлении достаточно велик (особенно для крупных институтов), и ученый секретарь с ней не справляется физически. В то же время библиотечные работники с современной подготовкой могли бы выполнять эти обязанности, поскольку они им достаточно близки в профессиональном плане, а свойственная им скрупулезность обеспечит качественное их выполнение.

Таким образом, в современных условиях для библиотечной системы РАН продолжают существовать свои направления деятельности, работая в которых она вносит свой вклад в развитие академической науки.

Литература

1. *Бочарова Е. Н., Докторов Я. Я.* Автоматизированная система формирования и ведения тематико-типологических планов в практике комплектования ЦБС БЕН РАН // Информационное обеспечение науки. Новые технологии: Сб. научн. тр. — М.: Научный мир, 2009. С. 200–207.
2. *Бочарова Е. Н., Кочукова Е. В., Докторов Я. Я.* Актуализация сводного тематико-типологического плана комплектования ЦБС БЕН РАН // Библиосфера, 2009. № 2. С. 87–89.
3. *Власова С. А., Докторов Я. Я., Калёнов Н. Е., Кочукова Е. В., Павлова О. В.* Интернет-технологии в развитии системы комплектования ЦБС БЕН РАН // Наукові праці Національної бібліотеки України імені В. І. Вернадського, 2010. Вып. 28. С. 41–55.
4. *Варакин В. П., Каленов Н. Е.* Управление ресурсами централизованной библиотечной системы // Информационные ресурсы России, 2010. № 115. С. 2–11.
5. *Варакин В. П., Власова С. А., Каленов Н. Е.* Современные информационные технологии в задачах обслуживания читателей ЦБС БЕН РАН // Вклад информационно-библиотечной системы РАН в развитие отечественного библиотековедения, информатики и книговедения: Юбилейный научный сборник, посвященный 100-летию ИБС РАН. — Новосибирск, 2011. С. 187–203.
6. *Власова С. А., Васильчиков В. В., Калёнов Н. Е., Левнер М. В.* Использование экспертных оценок для комплектования централизованных библиотечных систем // Научно-техническая информация. Сер. 1, 2007. № 5. С. 22–26.
7. *Власова С. А., Кочукова Е. В.* Экспертная система ЦБС БЕН РАН // Библиотеки национальных академий наук: проблемы функционирования, тенденции развития: Научн.-практ. и теор. сб. — Киев: Наукова Думка, 2010. Вып. 8. С. 79–85.
8. *Armstrong K., Starratt J.* Even cowgirls get the blues: How research libraries are coping with reductions in their collections budgets // World Library and Information Congress: 77th IFLA General Conference and Assembly. Puerto Rico, 2011. http://conference.ifla.org/sites/default/files/files/papers/ifla77/113-armstrong-en.pdf.
9. *Глушановский А. В., Калёнов Н. Е., Лексикова Е. Е.* База данных «SCIENCE CITATION INDEX» на CD-ROM. — М.: Биоинформсервис, 1993. 38 с.
10. *Лаврентьева М. В., Мелконян М. К., Смирнов С. Н.* Характеризация некоторых научных направлений Института кристаллографии РАН в базе данных «Web of Science» // Новые технологии в информационном обеспечении науки: Сб. научн. тр. — М.: Биоинформсервис, 2001. С. 127–131.
11. *Елепов Б., Лаврик О., Свирюкова В.* К подсчету индексов готовы // Наука в Сибири, 2007. № 35(2620). С. 8.
12. *Трескова П. П.* Наука в информационном измерении: анализ публикационной активности ученых с использованием баз данных «Web of Science» и «SCOPUS» // Информационное обеспечение науки. Новые технологии: Сб. научн. тр. — М.: Научный мир, 2009. С. 253–262.
13. *Каленов Н. Е., Савин Г. И., Сотников А. Н.* Электронная библиотека «Научное наследие России»: технология наполнения // Новые технологии в информационном обеспечении науки: Сб. научн. тр. — М.: Научный мир, 2007. С. 40–48.
14. *Каленов Н. Е., Савин Г. И., Сотников А. Н.* Электронная библиотека «Научное наследие России» как составляющая интеграционных процессов // Вестник Библиотечной Ассамблеи Евразии, 2011. № 3. С. 52–55.
15. *Бездушный А. Н, Бездушный А. А., Нестеренко А. К., Серебряков В. А., Сысоев Т. М., Теймуразов К. Б., Филиппов В. И.* Информационная Web-система «Научный институт на платформе ЕНИП». — М.: ВЦ РАН, 2007.

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2012. Т. 6. Вып. 2. С. 59–76

УНИФИКАЦИЯ ЯЗЫКОВ СИСТЕМ НА ПРАВИЛАХ ДЛЯ ОБЕСПЕЧЕНИЯ ИНТЕРОПЕРАБЕЛЬНОСТИ ДЕКЛАРАТИВНЫХ ПРОГРАММ*

Л. А. Калиниченко[1], С. А. Ступников[2]

Аннотация: Проанализированы рекомендации W3C RIF (Rule Interchange Format), ориентированные на обеспечение интероперабельности разнообразных систем на правилах введением расширяемого семейства унифицированных языков (диалектов) на правилах, позволяющих создавать сохраняющие семантику отображения в диалекты языков различных систем на правилах. Для определения мотивации проекта RIF дан краткий обзор развития и применения языков и систем на правилах в областях представления знаний, дедуктивных баз данных, логических моделей рассуждений. Также проанализированы различные семантики логических языков на правилах, оказавших влияние на конструкцию RIF. Рассмотрены основные классы применений интероперабельных программ на правилах, на основе которых были выработаны требования к RIF. Рассмотрены основные решения, принятые в проекте RIF.

Ключевые слова: унификация языков; расширяемость языков; системы логического программирования; системы на активных правилах; продукционные системы; представление знаний; дедуктивные базы данных; логические модели рассуждений; стратифицированная семантика; стабильная модель логической программы; хорошо обоснованная семантика; диалекты RIF; каркас RIF

1 Введение

Многообразие информационных технологий (ИТ) и их воплощений в конкретных ИТ-продуктах проявляется в многообразии языков, предназначенных для спецификации предметных областей, программ, схем баз данных, онтологий, интерфейсов информационных ресурсов (ИР), реализованных в рамках определенной ИТ, и пр. Число таких языков, их разнообразие со временем быстро растет, порождая сложные проблемы интеграции и интероперабельности разноязыких ИР. Стандартизация языков несколько ограничивает разнообразие, однако число стандартов остается большим, а конкретные реализации одного и того же стандарта зачастую остаются несовместимыми.

Примерами классов подобных языков являются языки реляционных и объектных баз данных, языки онтологического моделирования, языки представления слабо структурированных, графовых, мультимедийных данных, языки представления баз знаний, языки логического программирования, дедуктивные языки запросов к базам данных, языки спецификации процессов (потоков работ), языки спецификации интерфейсов программных ИР для обеспечения их (ИР) интероперабельности, языки со специализированной семантикой (например, для выражения темпоральных, пространственных моделей), языки для определения нечетких, вероятностных представлений, языки концептуального моделирования и метамодели и многие другие.

Проблемы унификации языков с целью нивелирования различий их синтаксиса и семантики исследуются давно. Так, язык IDL был разработан OMG в качестве стандарта языка спецификации интерфейсов объектов или программных компонентов, полученных в результате применения различных языков и систем программирования. Язык YAWL, основанный на сетях Петри, ориентирован на унифицированное представление разнообразных образцов поведения, представимых на различных языках описания процессов. Язык TSQL ориентирован на представление темпоральных реляционных баз данных.

Среди подходов к унификации языков особое место занимает концепция расширяемых языков, в которых фиксируется ядро, позволяющее унифицировать некоторую совокупность простых языков в определенном классе, и над таким ядром

*Настоящая публикация является журнальным вариантом текста доклада Л. А. Калиниченко и С. А. Ступникова «Анализ мотивации, целей и подходов проекта унификации языков на правилах», опубликованного в сборнике трудов Второго симпозума «Онтологическое моделирование», М.: ИПИ РАН, 2011. Работа выполнена при финансовой поддержке РФФИ (проекты 08-07-00157-а, 10-07-00342-а, 11-07-00402-а) и Программы фундаментальных исследований Президиума РАН № 15, проект 4.2.

[1]Институт проблем информатики Российской академии наук, leonidk@synth.ipi.ac.ru

[2]Институт проблем информатики Российской академии наук, ssa@ipi.ac.ru

надстраиваются расширения, каждое из которых вместе с ядром является результатом сохраняющего семантику отображения некоторого исходного языка (языков) [1]. Развитым проектом расширяемого языка является язык СИНТЕЗ [2, 3], сопровождаемый методами и средствами построения его расширений. Основным средством поддержки процесса построения расширений языка-ядра является Унификатор информационных моделей [4], позволяющий конструировать расширения языка СИНТЕЗ и отображения конкретных исходных языков в такие расширения. При этом сохранение семантики операторов трансформации состояний и поведения в исходном языке в их отображении в язык СИНТЕЗ основано на принципе уточнения[1] [5]. Таким образом, можно конструировать доказательно правильные расширения языка и отображения конкретных языков в язык СИНТЕЗ. Эти методы применялись к отображению в язык СИНТЕЗ языков баз данных, процессных языков, онтологических языков, языков спецификации объектных интероперабельных компонентов и др.

Вместе с тем, за исключением языков манипулирования данными в базах данных, вопросы унификации языков программирования не рассматривались вообще. Поэтому инициатива W3C, известная как RIF, заслуживает в контексте работ по унификации языков особого внимания [6, 7]. Использование и развитие языков на правилах для логического программирования, дедуктивных баз данных, представления знаний, создания интеллектуальных информационных систем продолжается уже более 30 лет. Технология языков на правилах созревала в этот период в результате теоретических исследований, практического и коммерческого применения таких языков. В частности, накопленный в этой области опыт существенно превосходит опыт в области дескриптивных логик, активно развиваемых для онтологического моделирования.

Для эффективного использования потенциала подходов, основанных на правилах, сообщества применения языков на правилах в области искусственного интеллекта, в бизнесе, в Семантическом Вебе организовали проект по выработке решений по унифицированному, интероперабельному использованию спецификаций, представленных на различных языках на правилах (таких языках, как в работе [8]). Применение результатов этого проекта должно быть общеупотребительным и не ограничиваться Семантическим Вебом. Идея RIF состоит в создании расширяемого унифицированного языка на правилах, обеспечивающего возможность построения сохраняющих семантику отображений в него различных языков на правилах. Такой унифицированный язык представляется как семейство диалектов, которые имеют общее ядро (корневой диалект) и совокупность расширяющих его диалектов, образующих ориентированный граф без циклов. Каждое ребро графа (отношение расширения диалектов) отправляется от более простого к расширенному диалекту. Для сравнения, при построении аксиоматических расширений языка СИНТЕЗ отношение расширения интерпретировалось как включение множества аксиом более простого языка во множество аксиом более сложного языка [1]. Итак, в RIF унифицированный язык на правилах представляется как семейство унифицированных языков (диалектов).

В RIF по отношению к диалекту каждая система программирования (система вывода) на правилах может выступать в двух независимых ролях — роли поставщика и роли потребителя. Первая роль означает, что система на правилах обеспечивает преобразование собственных программ в программы на унифицированном диалекте. Вторая роль означает, что система на правилах может воспринимать программы на диалекте и преобразовывать их в программы на собственном языке системы. Таким образом, для полновесного включения каждого языка на правилах в совокупность интероперабельных языков достаточно снабдить соответствующую систему программирования двумя сохраняющими семантику преобразователями — из собственного языка в адекватный диалект (роль поставщика) и из диалекта в собственный язык (роль потребителя).

Для работы по проекту RIF в 2005 г. была образована рабочая группа W3C RIF WG [9] с целью выработки «обменного формата» правил — так кратко именуется семейство определяемых RIF унифицированных диалектов языка на правилах.

Изначально было декларировано, что целью RIF не является обеспечение единственного языка, объемлющего черты всех известных языков на правилах. Различные средства различных языков и систем на правилах зачастую не совместимы. Поэтому и была принята концепция диалектов RIF, обеспечивающая возможность обмена модулями правил между различными системами на правилах. По замыслу проекта для правил, созданных в рамках некоторого приложения, должна быть обеспечена возможность их публикации, использования совместно с другими правилами, повторного использования в других приложениях и других системах на правилах.

[1] Говорят, что спецификация A уточняет спецификацию B, если систему, удовлетворяющую A, можно использовать вместо системы, удовлетворяющей B, и при этом пользователь не замечает этой замены.

Кроме того, важной задачей RIF WG было обеспечение возможности совместной работы диалектов RIF и существующих стандартов Семантического Веба — таких как RDF (Resource Description Framework) и OWL (Web Ontology Language), не совместимых с большинством существующих языков на правилах. В частности, важно было показать, что существует возможность совместной работы дескриптивных логик и систем на правилах:

- сформировать общий базис для онтологических языков на правилах и на дескриптивной логике с целью достижения их интероперабельности;
- проанализировать пересечение этих двух формализмов представления знаний для понимания того, какой выразительности можно достичь при их объединении (комбинации);
- обеспечить использование машин на правилах как масштабируемых служб рассуждений в онтологиях, определяемых на таком пересечении.

В октябре 2009 г. был завершен важный этап RIF — публикация документов спецификации RIF [7] в качестве кандидатов для стандартизации W3C:

(1) RIF Overview;
(2) RIF Core Dialect;
(3) RIF Basic Logic Dialect;
(4) RIF Framework for Logic Dialects;
(5) RIF RDF and OWL Compatibility;
(6) RIF Datatypes and Built-Ins 1.0;
(7) RIF Production Rule Dialect;
(8) RIF Test Cases;
(9) RIF Combination with XML data;
(10) OWL 2 RL in RIF.

В июне 2010 г. документы 2–6 были приняты в качестве стандарта W3C. Создание общепринятой спецификации RIF — сложная задача. Выполненная RIF WG работа впечатляет глубиной замысла, масштабом, точностью разработанных, тщательно формализованных спецификаций. Можно ожидать, что предложенная концепция окажет существенное влияние на развитие декларативных языков и их применение в различных областях, расширяя тем самым существующие пределы использования систем на правилах. Результаты, полученные RIF WG, заслуживают детального анализа и изучения. Настоящая работа посвящена именно такому анализу.

Статья организована следующим образом. В разд. 2 дан краткий анализ истории развития и применения языков и систем на правилах, включая области представления знаний, дедуктивных баз данных, логических моделей рассуждений. В разд. 3 приведен обзор различных семантик логических программ на правилах, таких как стратифицированная семантика, семантика стабильной модели логической программы, хорошо обоснованная семантика. В разд. 4 рассмотрены основные классы применений интероперабельных программ на правилах, на основе которых были выработаны требования к RIF. В разд. 5 рассмотрены основные решения, принятые в проекте RIF.

2 Языки на правилах и их применение: краткий исторический экскурс

Настоящий раздел является введением в контекст, в рамках которого формировался проект RIF. При написании раздела использовались материалы известных обзоров симбиотического развития логики и программирования, логики и баз данных, а также применения полученных результатов для представления знаний и программирования различных моделей рассуждений в приложениях [10–24].

2.1 Программы на правилах, классы применений

Логическое программирование характеризуется тремя основными классами применений: в качестве универсального языка программирования, языка баз данных, языка представления знаний. В качестве языка программирования оно позволяет представить и вычислить любую вычислимую функцию. Как язык баз данных оно обобщает реляционные базы данных, позволяя наряду с фактами представлять правила общего вида. Наконец, как язык представления знаний оно является немонотонной логикой, которую можно использовать для рассуждений по умолчанию (подробнее о моделях рассуждений см. п. 2.4).

Логические программы представляют собой множества условных высказываний вида:

if B_1 *and* . . . *and* B_n *then* H ,

в которых *следствие* H представляет собой атомарную формулу, а *условия* B_i являются *литералами*, представляющими собой атомарные формулы или их отрицания. Все переменные неявно связаны квантором всеобщности, располагаемым перед условным высказыванием. Условные высказывания в логических программах называются *клаузами*. Факты являются клаузами специального вида,

в которых $n = 0$ (нет условий) и нет переменных. Клаузы, не являющиеся фактами, называются *правилами*. Цели (или *запросы*) представляют собой конъюнкции литералов, подобно условиям в клаузах. Однако все переменные в них неявно связаны квантором существования, и задачей цели является отыскание означиваний переменных цели, при которых цель приобретает истинное значение.

При *обратных рассуждениях* (от следствий к условиям) условные высказывания рассматриваются как процедуры редуцирования цели: чтобы показать, что H, нужно разрешить B_1 *and* . . . *and* B_n.

Поскольку условные высказывания в логических программах рассматриваются в таком обратном порядке, обычно они так и записываются:

H *if* B_1 *and* . . . *and* B_n,

так что обратные рассуждения становятся эквивалентными «прямому связыванию» или «прямому сцеплению» в направлении, в котором записывается условное высказывание. В синтаксисе языка Prolog клауза выглядит так:

H: - $B_1, \ldots, B_n$.

В таком виде клаузы можно трактовать либо декларативно как условные высказывания в обратной записи, либо процедурно как процедуры редуцирования цели, исполняемые в прямом направлении.

Позитивные атомарные цели и подцели разрешаются в процессе обратных рассуждений. Цели и подцели с отрицанием вида *not* G, где G — атомарное предложение, разрешаются при трактовке отрицания как неудачи, согласно которой *not* G удовлетворяется тогда и только тогда, когда обратные рассуждения с подцелью G терпят неудачу. Отрицание как неудача превращает логическое программирование в немонотонную логику.

Цели и условия клауз могут быть обобщены, так что вместо конъюнкций литералов могут быть использованы произвольные формулы логики первого порядка.

2.2 Применение логики для представления знаний и решения задач

В этом пункте кратко рассмотрено развитие подходов, основанных на применении языков на правилах для представления знаний и решения задач [22]. Рассматриваемые подходы важны для понимания мотивации проекта RIF.

К концу 1960-х гг. в подходах к искусственному интеллекту наметились две основных тенденции: эвристический подход, связываемый главным образом с продукционными системами и стремлением к представлению специальных знаний конкретных предметных областей, и формальный подход, применявшийся в системах, основанных на принципе резолюции, в рамках которого подчеркивалась важность универсальных методов решения задач, не зависящих от предметной области.

Процедурный подход как альтернатива логическому находит выражение в системах Planner и micro-Planner (MIT). В 1971 г. Виноградом на основе системы micro-Planner был реализован диалог на естественном языке для простой предметной области. Ковальски в сотрудничестве с Кольмероэ предпринял попытку повторить в логике реализацию системы Винограда на основе принципа резолюции. Это привело к реализации процедурной интерпретации Хорновских клауз в 1974 г. и к разработке языка программирования Prolog в начале 1970-х гг. в Марсельском университете А. Кольмероэ и его сотрудниками на основе теоретических работ Р. Ковальски (программа доказательства теорем, которая включала интерпретатор Хорновских клауз Ковальски).

В середине 1970-х гг. Марвин Минский (MIT) предпринял очередное наступление против логического подхода, предложив фреймы для представления знаний по умолчанию без необходимости строгой и точной спецификации исключений.

Логическое сообщество ответило немонотонными логиками, включая очерчивание МакКарти (1980 г.), логику умолчания Рейтера (1980 г.), аутоэпистемическую логику Мура (1985 г.), обоснованное Кларком отрицание как неудачу в логическом программировании (1978 г.).

В этот же период было замечено, что рассуждение по умолчанию можно интерпретировать как форму абдуктивных рассуждений. Основываясь на этих результатах, Ковальски показал, что отрицание как неудачу в логическом программировании можно также интерпретировать в таких терминах. Дунг (2006 г.) показал, что большинство немонотонных логик может быть интерпретировано в терминах логики аргументации на основе абдуктивных гипотез.

Абдуктивные модели логического программирования (Abductive Logic Programming, ALP) используют специальную трактовку различий между данными и ограничениями целостности в базах данных для интерпретации убеждений как данных и целей как ограничений целостности. В традиционных базах данных ограничения целостности используются пассивно, чтобы воспрепятствовать неправильным изменениям данных. В абдуктивном логическом программировании они играют роль, подобную ограничениям целостности в активных базах данных, в которых они, с одной стороны,

препятствуют неправильным изменениям, а с другой — выполняют корректирующие действия для поддержания целостности базы данных.

Продукционные правила, используемые как ассоциации стимул–реакция, могут рассматриваться как ограничения целостности такого вида. По сравнению с классической логикой, которой соответствует как декларативная теоретико-модельная семантика, так и различные процедуры доказательства, продукционные системы не имеют декларативной семантики вообще. Продукционные правила имеют вид логических импликаций, не обладая семантикой импликаций.

Область дедуктивных баз данных [19], в особенности область обработки рекурсивных запросов, становится быстро развивающейся во второй половине 1980-х – начале 1990-х гг. Дедуктивные базы данных заимствуют значительное число концепций от логического программирования. Так, правила и факты, представляемые в языке дедуктивных баз данных Datalog, похожи на представления программ на языке Prolog. Вместе с тем существует ряд существенных различий между дедуктивными базами данных и логическим программированием. Одним из основных различий является то, что логические программы ориентированы на покортежную обработку, в то время как дедуктивные базы данных оперируют множествами. Языки различных систем дедуктивных баз данных отличаются видом поддерживаемой рекурсии, интерпретацией отрицания, способностью поддерживать ограничения целостности, видом поддерживаемых внешних интерфейсов.

Если первоначально в работах по языкам логического программирования преобладало стремление достижения декларативности программ, то в дальнейшем доминирующим стало намерение поддержать средствами языков различные модели логических рассуждений. В большинстве своем системы поддержки логических рассуждений являются немонотонными (например, в области принятия решений, обработки исключительных ситуаций, аргументации, байесовских стратегий, статистического вывода и пр.). Существенным для логических рассуждений моментом является трактовка ими предположений о замкнутости или открытости мира, а также интерпретация отрицания (например, отрицание как неудача [10, 15, 17, 21] связано с предположением о замкнутости мира, поскольку согласно ему каждый предикат считается ложным, если нельзя доказать его истинность).

К подобным моделям логических рассуждений относятся, например, следующие.

Рассуждения на основе умолчания [12, 16, 23]. Этот вид рассуждений поддерживается логиками умолчания [16], отмены заключений, аргументации, программирования множества ответов (Answer Set Programming, ASP).

Абдуктивные рассуждения [24] составляют процесс получения наиболее вероятных объяснений известных фактов. Абдуктивная логика является немонотонной, поскольку наиболее вероятные объяснения не обязательно являются правильными. Абдукция — познавательная процедура принятия гипотез.

Логика очерчивания (*circumscription*) — это немонотонная логика, созданная Дж. МакКарти для формализации предположений здравого смысла, которые должны действовать по отношению к сущностям реального мира, если нет других указаний.

Логика отмены заключений (*defeasible reasoning*) — это разновидность рассуждений, основанных на суждениях, которые являются отменяемыми, в отличие от неотменяемых суждений, характерных для дедуктивной логики (такая способность является важной, например, при принятии решений).

Логика рассуждений о действиях и изменениях. Большая часть работ в этой области посвящена применению ситуационного исчисления — формализма, предложенного Дж. МакКарти для описания действий, рассуждений о них и эффектов действий. Активно исследуются логики действий, применение модальных логик для рассуждений о знаниях и действиях.

Логика рассуждений с неопределенностью. Сюда относятся статистические методы обнаружения закономерностей в данных.

Следует заметить, что каждая из перечисленных моделей логических рассуждений является предметом серьезных исследований, результаты которых находят применение в различных областях. Эти исследования тесно связаны с развитием языков логического программирования с адекватной семантикой.

3 Разнообразие семантик логических программ

3.1 Минимальная модель программы. Стратифицированная семантика

Клауза Хорна представляет собой конъюнкцию литералов — атомарных формул (атомов) или их отрицаний с не более чем одним атомом в голове. Клауза Хорна с одним атомом в голове называется *определенной* (definite) клаузой. Клаузы Хорна играют основополагающую роль в логическом программировании. Клауза Хорна, не содержащая атома в голове, называется целевой клаузой. Резолюция целевой клаузы с определенной клаузой, при которой

образуется новая целевая клауза, составляет основу SLD-резолюции (Selective Linear resolution with Definite clauses) — одного из способов интерпретации логических программ. Теоретико-модельная семантика подразумевается при реализации вывода в прямом направлении (снизу вверх) или в обратном направлении (сверху вниз)[1].

Семантика логической программы определяется ее минимальной моделью. Программа, содержащая только атомы без отрицаний, называется позитивной. Каждой позитивной программе соответствует единственная минимальная модель, называемая наименьшей моделью. Определение наименьшей модели достигается оператором вычисления наименьшей неподвижной точки логической программы.

Позитивные логические программы позволяют реализовать декларативное моделирование при решении разнообразных задач. Во многих случаях, однако, требуется использование отрицания, нуждающегося в адекватной семантической интерпретации. Семантика языка Prolog представляет собой развитие от SLD к SLDNF резолюции на основе идеи отрицания как неудачи (Negation As Failure, NAF).

Реализация отрицания как неудачи в языке Prolog является проблематичной [25]: использование NAF в теле клауз приводит к сложностям в случае рекурсии при наличии циклических зависимостей предикатов (атомов) в правилах; поэтому современные системы логического программирования используют либо хорошо обоснованное отрицание по умолчанию [26], либо отрицание с семантикой, определяемой стабильной моделью [15, 27].

Важный класс программ с отрицанием составляют *стратифицированные программы*. Они обладают тем свойством, что можно установить порядок вычисления правил программы, при котором значения атомов с отрицанием могут быть предопределены. Иными словами, для вычисления тела правила, содержащего *not* $r(t)$, значение атома с отрицанием $r(t)$ должно быть определено. С этой целью предикаты вычисляются по слоям (стратам) программы снизу вверх. Этот подход работает, если в программе не возникает циклов с предикатами, содержащими отрицание.

Стратификация заключается в любом непротиворечивом присваивании номеров символам предикатов, гарантирующем существование однозначной формальной интерпретации логической программы. Говорят, что набор клауз вида

$$Q_1 \wedge \cdots \wedge Q_n \wedge \neg\ Q_{n+1} \wedge \cdots \wedge \neg\, Q_{n+m} \rightarrow P$$

стратифицирован тогда и только тогда, когда существует стратификационная нумерация, удовлетворяющая следующим условиям:

1. Если предикат P позитивно выводим из предиката Q (т. е. P находится в голове правила, а Q позитивно входит в тело этого же правила), то стратификационный номер P должен быть бóльшим или равным стратификационному номеру Q: $S(P) \geq S(Q)$.

2. Если предикат P выводим из предиката с отрицанием Q (т. е. P находится в голове правила, а Q входит с отрицанием в тело этого же правила), то стратификационный номер P должен быть бóльшим стратификационного номера Q: $S(P) > S(Q)$.

Понятие стратифицированного отрицания позволяет получить эффективную операционную семантику стратифицированной программы на основе стратифицированной наименьшей неподвижной точки, вычисляемой итеративно. Оператор получения неподвижной точки применяется к каждой страте программы, двигаясь от страт с меньшими номерами к стратам с большими номерами.

В контексте IDB (Intensional DataBase) и EDB (Extensional DataBase) последняя, представляющая собой набор фактов, получает номер (ранг) 0. Предикаты IDB, правила определения которых не включают отрицаний, также имеют ранг 0. Предикаты IDB, чьи единственные отрицательные зависимости выражаются посредством предикатов ранга 0, получают ранг 1 и т. д. Стратифицируемость легко установить синтаксически анализом одной лишь IDB.

Определение стратифицированной семантики реализуется индуктивно. После того как все атомы с рангом, меньшим k, были классифицированы как позитивные или атомы с отрицанием, эти литералы используются для получения значений позитивных атомов ранга k и определения $\neg q$ для всех атомов q ранга k, которые не были выведены. Полученный результат называется стратифицированной моделью.

Стратифицированная семантика согласуется с семантикой хорошо обоснованной модели для всех рангов. Нестратифицируемым программам соответствует хорошо обоснованная семантика или семантика стабильной модели.

[1] При обратном рассуждении атомарный запрос унифицируется с головой правила и заменяется соответствующим экземпляром тела. При прямом рассуждении голова означенного экземпляра правила включается в множество следствий после того, как все атомы тела этого правила уже были включены в множество следствий.

3.2 Семантика стабильной модели логической программы

В логическом программировании существует два принципиально разных взгляда на семантику программ, отражающих две философски различные точки зрения [28].

Первый подход отражает стремление сохранить единственную модель программы даже для проблемных классов программ с отрицанием. Этого можно достичь, определяя надлежащим образом выбор единственной модели среди всех возможных моделей программы. Наиболее популярная семантика в этом подходе основана на *хорошо обоснованной модели*.

Второй подход характеризует противоположная точка зрения — соотнесение программе множества моделей, отбрасывая «догматическое» требование единственности модели. В общем случае считается, что одной программе может соответствовать множество совместимых с ней сценариев получения модели. В рамках этого подхода говорят о генерации множества *стабильных моделей*. Этот подход выходит за рамки простого ответа на запрос, речь идет о получении решения задач.

Далее оба подхода рассматриваются более подробно, начиная с семантики стабильной модели. Интуитивно семантика стабильной модели основана на особой трактовке атомов с отрицанием, являющихся источником «противоречий» или «нестабильности». «Стабильность» при этом заключается в следующем. Если интерпретация M программы P непротиворечива, то она стабильна.

Простой пример программы, которой соответствует множество моделей:

man(*petrov*).
single(X): - *man*(X), *not husband*(X).
husband(X): - *man*(X), *not single*(X).

Здесь утверждения *single*(*petrov*) и *husband*(*petrov*) взаимозависимы при использовании отрицания. Алгоритм SLD-резолюции зациклился бы здесь при попытке ответить на запрос *single*(X). Вместе с тем у этой программы есть две минимальных модели Эрбрана, являющихся стабильными:

$M_1 = \{man(petrov), single(petrov)\}$,
$M_2 = \{man(petrov), husband(petrov)\}$.

Подход к логическому программированию на основе семантики стабильной модели [27], выражающий применение идеи аутоэпистемической логики [29] и логики умолчания [10] к анализу отрицания как неудачи, называется *программированием множества ответов* (ASP) [28]. Возможность использования средств вывода множества ответов как новой парадигмы программирования была обоснована в [30] (название «программирование множества ответов» было впервые употреблено в заголовке соответствующей части сборника, включающего эту статью) и в [31].

Язык ASP отличается от многих языков представления знаний способностью выражать утверждения, основанные на умолчании, вида «обычно экземпляры класса C удовлетворяют свойству P». Выражение умолчаний, исключений из таких умолчаний, а также способов использования этой информации для вывода адекватных заключений может быть поддержано ASP. Язык ASP позволяет также выражать причинный эффект действий («утверждение F становится истинным в результате выполнения действия A»), утверждений, выражающих недостаток информации («неизвестно, является ли утверждение P истинным или ложным»), различные предположения общего вида, например «утверждения, не следующие из базы знаний, являются ложными».

В ASP как в язык программирования синтаксически добавляются дизъюнкции, сильные отрицания (наряду со слабым отрицанием, использующем семантику отрицания как неудачи), ограничения.

В дизъюнктивном правиле голова может представлять собой дизъюнкцию нескольких атомов:

$$A_1 \vee \ldots \vee A_k \text{: - } B_1, \ldots, B_m, \mathit{not}\ C_1, \ldots, \mathit{not}\ C_n.$$

Например, можно использовать правило вида

$$\mathit{female}(X) \vee \mathit{male}(X) \text{: - } \mathit{person}(X).$$

Пример записи дизъюнктивного факта:

$$\mathit{broken}(\mathit{left\_hand}, \mathit{tom}) \vee \mathit{broken}(\mathit{right\_hand}, \mathit{tom}).$$

Правило

$$\mathit{ok}(C) \vee \neg\ \mathit{ok}(C) \text{: - } \mathit{component}(c)\ .$$

утверждает, что компонент может находиться в рабочем состоянии или не работать.

Для выражения правил с умолчанием используется сильное отрицание в комбинации со слабым. Например, выражение того, что «по умолчанию птица летает», обеспечивается правилом

$$\mathit{flies}(X) \text{: - } \mathit{bird}(X), \mathit{not}\ \neg\ \mathit{flies}(X).$$

Здесь $\neg$ — сильное отрицание. Предикат *not* p в теле правила можно интерпретировать (согласно В. Лифшицу) как «не верится, что p».

Программы в ASP могут включать также правило выбора, такое как

$$\{s, t\} \text{: - } p\ .$$

Это правило означает следующее. Если p включается в стабильную модель, то следует выбрать произвольным образом, какой из атомов — s или t — нужно включить.

Простое расширение позволяет использовать в программах ограничения правила, в которых отсутствует голова:

$$: \text{-}\ B_1, \ldots, B_m, \mathit{not}\ C_1, \ldots, \mathit{not}\ C_n.$$

Здесь $B_1, \ldots, B_m, C_1, \ldots, C_n$ — атомы. Ограничению соответствует отрицание формулы, эквивалентной его телу:

$$\neg\,(B_1 \wedge \ldots \wedge B_m \wedge \neg\, C_1 \wedge \ldots \wedge \neg\, C_n).$$

В ASP ограничения играют важную роль. Включение ограничения в логическую программу P оказывает влияние на набор стабильных моделей P: стабильные модели, нарушающие ограничение, исключаются из такого набора. Иными словами, для любой логической программы P с ограничениями и любого ограничения C стабильные модели $P \cup \{C\}$ включают те стабильные модели P, которые удовлетворяют C.

Далее следует пример спецификации задачи трехцветной (*b*, *r*, *g*) раскраски графа $G = (V, E)$, представленного узлами *node*(*n*) для каждого $n \in V$ и ребрами *edge*(*n*, *n′*) для каждой пары $(n, n') \in E$, задаваемой следующими правилами:

b(*X*): - *node*(*X*), *not r*(*X*), *not g*(*X*).
r(*X*): - *node*(*X*), *not b*(*X*), *not g*(*X*).
g(*X*): - *node*(*X*), *not r*(*X*), *not b*(*X*).

и ограничениями

: - *b*(*X*), *b*(*Y*), *edge*(*X*, *Y*).
: - *r*(*X*), *r*(*Y*), *edge*(*X*, *Y*).
: - *g*(*X*), *g*(*Y*), *edge*(*X*,*Y*).

Логическое программирование в парадигме ASP успешно применяется для решения задач в широком классе проблем, включая диагностику в разных областях, интеграцию информации, поиск решения при задании набора ограничений, планирование действий, прокладку маршрутов, проблемы биомедицины и биологии, извлечение информации из текстов, классификацию.

Пусть P представляет собой множество правил вида

$$A\text{: -}\ B_1, B_2, \ldots, B_m, C_1, C_2, \ldots, C_n.$$

Здесь A, B_1, B_2, $\ldots$, B_m, C_1, C_2, $\ldots$, C_n — базовые (ground) атомы. Если P не содержит отрицаний $(n = 0)$ в каждом правиле программы, то по определению единственной стабильной моделью P является ее минимальная модель. Чтобы расширить это определение на случай программ с отрицанием, понадобится вспомогательное понятие редукта, определяемого следующим образом.

Для любого множество I базовых атомов *редуктом* P по отношению к I называется множество правил без отрицания, полученных из P исключением каждого правила, такого что, по крайней мере, один из атомов C_i в теле этого правила принадлежит I, а затем исключением частей *not* C_1, *not* C_2, $\ldots$, *not* C_n из тел оставшихся правил.

Это *преобразование Гельфонда–Лифшица* [32], иногда называемое стабильным преобразованием. Целью такого преобразования является представление моделей как множества базовых атомов, в котором отсутствующие атомы представляют атомы с отрицанием. В этом контексте «минимальная модель» — это модель, содержащая минимальное множество позитивных атомов, а «монотонность преобразования» полных интерпретаций заключается в том, что оно монотонно при рассмотрении одних лишь позитивных атомов. Стабильные модели представляются в двузначной логике.

Говорят, что I — это *стабильная модель* P, если I — стабильная модель редукта P по отношению к I. Каждая стабильная модель P является моделью P. Полная модель логической программы P является стабильной, если она является неподвижной точкой трансформации Гельфонда–Лифшица. Если программа P имеет в точности одну стабильную модель, то она называется уникальной стабильной моделью P.

3.3 Хорошо обоснованная семантика и ее соотношение с ASP

Хорошо обоснованной семантикой Web Feature Service (WFS) программы P согласно определению из [26] является ее значение, представленное наименьшей неподвижной точкой трансформации множества литералов, атомы которых входят в базу Эрбрана данной программы P. Каждый позитивный литерал означает, что его атом является истинным, каждый литерал с отрицанием означает, что его атом является ложным, а атомы не имеют присвоенного им значения истинности. Таким образом, эта модель является моделью трехзначной логики.

Например, если известно, что

Объект A — это ночная бабочка,
если A не летает днем,

но неизвестно, летает ли *A* днем, в хорошо обоснованной семантике высказывание «объект *A* — это ночная бабочка» получает значение *unknown*, т. е. его значением не является ни истина, ни ложь.

Если атом является истинным в хорошо обоснованной модели программ P, то он принадлежит каждой стабильной модели P. Обратное утверждение, вообще говоря, не выполняется. Например, программа

p: - *not q*.
q: - *not p*.
r: - *p*.
r: - *q*.

имеет две стабильные модели $\{p, r\}$ и $\{q, r\}$. Несмотря на то что *r* принадлежит обеим моделям, значением *r* в хорошо обоснованной модели является *unknown*.

Более того, если атом является ложным в хорошо обоснованной модели программы, то он не принадлежит ни к одной ее стабильной модели. Таким образом, хорошо обоснованная модель логической программы является нижней гранью пересечения ее стабильных моделей и верхней гранью их объединения.

Одним из основных отличий ASP от хорошо обоснованной семантики является ориентация ASP на решение переборных задач, что невозможно в WFS. Из-за различия выразительной способности двух парадигм они используются в разных целях. Язык ASP идеально подходит для решения сложных комбинаторных задач, соответствующие этой парадигме системы обычно применяются как компоненты поддержки баз знаний, встраиваемые в императивные системы программирования. Например, система LPARSE первоначально была создана как фронтальный процессор для решателя множества ответов SMODELS, а впоследствии использовалась аналогичным образом с большинством других решателей множества ответов. Система DLV [33] является исключением: синтаксис ASP-программ в DLV отличается от синтаксиса, используемого в других системах.

Диалекты логического программирования, такие как Datalog с немонотонным отрицанием, над которыми надстраивается ASP (как расширение Datalog'a), часто рассматриваются в качестве естественного основания для слоя правил Семантического Веба. Современные системы ASP содержат расширения для извлечения данных в RDF и задания запросов к OWL. В архитектуре Семантического Веба изучаются проблемы, возникающие в связи с добавлением правил с немонотонным отрицанием, основанным на предположении о замкнутости мира, над RDF и OWL, основанных на предположении об открытости мира[1].

В противоположность ASP, WFS-базированные системы являются вычислительно полными и используются как полновесные системы программирования. Хорошо обоснованная семантика является основой реализации многих систем (например, XSB, Ontobroker, Intellidimension, SweetRules, SILK, FLORA).

3.4 Фреймовая логика и язык метапрограммирования

В этом пункте кратко рассматриваются элементы языков F-Logic и HiLog (в их воплощении в языке системы на правилах FLORA-2 [34]), которые оказали существенное влияние на формирование каркаса RIF.

Фреймовая логика (F-logic) [35, 36] рассматривается как язык представления знаний и как онтологический язык. Она соединяет возможности концептуального моделирования с объектно-ориентированными, фреймово-базированными языковыми средствами и предлагает компактное, декларативное представление программ в хорошо обоснованной семантике логического языка программирования. Язык поддерживает такие средства, как, например, уникальная идентифицируемость сложно структурированных объектов, наследование, полиморфизм, методы, поддерживающие запросы, инкапсуляция. Первоначально F-logic как язык был разработан в ориентации на дедуктивные базы данных, но в последнее время он используется все чаще для поддержки семантических технологий. При этом F-logic обеспечивает логические основания фреймово-базированного и объектно-ориентированного языка представления данных и знаний.

HiLog — это логическое расширение языка Prolog для поддержки средств метапрограммирования и программирования в логике высоких порядков при сохранении вычислимости в логике первого порядка.

Язык FLORA-2 — это диалект F-logic с многочисленными расширениями, включая метапрограммирование в стиле языка HiLog и возможности изменения баз данных логическими средствами, представленными в виде *транзакционной логики*. FLORA-2 обладает также развитыми средствами проектирования модульного программного обеспечения на основе динамических модулей. Области применения FLORA-2 включают область интеллигентных агентов, Семантический Веб, сети баз знаний, онтологическую инженерию, интеграцию информации. В основе FLORA-2 используется машина вывода XSB.

[1] Эта проблема подробно изучается в проекте RIF.

Колорит языка логического программирования F-logic [34] придает ему немонотонную семантику при интерпретации операции отрицания как неудачи, поддержке множественного наследования с перекрытием, что выходит за рамки традиционного логического программирования.

В синтаксисе F-logic выражение принадлежности экземпляра классу выглядит как *John:student*, а выражение отношения подкласса реализуется как *student::person*. Сами классы интерпретируются как объекты, так что один и тот же объект может играть роль класса в одной формуле и роль объекта в другой. Например, в формуле *student:class* символ *student* играет роль объекта, тогда как в формуле *student::person* он выступает в роли класса.

F-logic также обеспечивает возможность задания информации о схеме при помощи сигнатурных формул. Например, *person*[*spouse* {0:1} ⇒ *person*, *name*{0:1} ⇒ *string*, *child* ⇒ *person*] — это сигнатурная формула, которая выражает тот факт, что класс *person* имеет три атрибута — однозначные атрибуты *spouse* и *name* (однозначность выражается при помощи ограничения кардинальности 0:1) и множественного атрибута *child*. При этом также сообщается, что значением первого атрибута являются объекты типа *person*, второй атрибут принимает значения типа *string*, а последний имеет значением множество, экземплярами которого являются объекты типа *person*.

Ниже представлен пример программы о базе данных публикаций на языке FLORA-2 [37]. Схема выглядит следующим образом:

paper[*authors* ⇒ *person*, *title* ⇒ *string*].
journal_p :: *paper*[*in vol* ⇒ *volume*].
conf_p :: *paper*[*at_conf* ⇒ *conf_proc*].
journal_vol[*of* ⇒ *journal*, *volume* ⇒ *integer*,
  number ⇒ *integer*, *year* ⇒ *integer*].
journal[*name* ⇒ *string*,
  publisher ⇒ *string*, *editors* ⇒ *person*].
conf_proc[*of_conf* ⇒ *conf_series*, *year* ⇒ integer,
  editors ⇒ *person*].
conf_series[*name* ⇒ *string*].
publisher[*name* ⇒ *string*].
person[*name* ⇒ *string*, *affil*(*integer*) ⇒ *institution*].
institution[*name* ⇒ *string*, *address* ⇒ *string*].

Частью программы являются также определения объектов:

*o_j*1: *journal_p*[*title* → '*Records*, *Relations*, *Sets*, *Entities*, *and Things*', *authors* → {*o_mes*}, *in vol* → *o_i*11].
o_di: *conf_p*[*title* → '*DIAM II and Levels of Abstraction*', *authors* → {*o_mes*, *o_eba*}, *at_conf* → *o_v76*].
*o_i*11: *journal_vol*[*of* → *o_is*, *number* → 1, *volume* → 1, *year* → 1975].
o_is: *journal*[*name* → '*Information Systems*', *editors* → {*o_mj*}].
o_v76: *conf_proc*[*of* → *vldb*, *year* → 1976, *editors* → {*o_pcl*, *o_ejn*}].
o_vldb: *conf_series*[*name* → '*Very Large Databases*'].
o_mes: *person*[*name* → '*Michael E. Senko*'].
o_mj: *person*[*name* → '*Matthias Jarke*', *affil*(1976) → *o_rwt*].
o_rwt: *institution*[*name* → '*RWTH Aachen*'].

Запросы можно задавать как относительно информации в схеме классов, так и относительно содержания структуры отдельных объектов. Это достигается помещением переменных в надлежащие синтаксические позиции. Так, запрос

? - *student*[?*M* ⇒ *person*]

находит методы, определенные на множествах в классе *student*, и возвращает объекты типа *person*. Следующий запрос возвращает все суперклассы класса *student*:

? - *student*::?*C and student*[*name* ⇒ ?*T*].

Для повышения гибкости таких метазапросов в HiLog введены синтаксические конструкции второго порядка, что позволяет использовать переменные на месте функциональных и предикатных символов. Например, допускаются запросы, подобные следующему:

? - *person*[?*M*(?*Arg*) ⇒ *integer*].

в котором переменная ?*M* связывается с функциями. Тем не менее семантика таких символов второго порядка остается первопорядковой. Это значит, что переменные связываются не с экстенсионалами символов (т. е. с отношениями, интерпретирующими символы предикатов или функций), а с самими такими символами.

HiLog допускает также использование переменных над атомарными формулами. Например, рассмотрим запрос, который завершается связыванием переменной ?*X* с атомом *p*(*a*):

p(*a*).
q(*p*(*a*)).
? - *q*(?*X*), ?*X*.

Здесь высказывание *p*(*a*) материализуется (is reified) в виде объекта, в результате чего оно может быть связано с переменной. Материализация атомарных формул в HiLog'е может быть расширена до произвольных бесквантворных формул языков HiLog и F-logic. Например, можно сказать, что *John* верит в то, что *Mary* нравится *Sally*: *John*[*believes* → $ {*Mary*[*likes* → *Sally*]}]. Обозначение ${. . . } в языке FLORA-2 используется для обозначения операторов материализации. Примером

более сложного оператора материализации является следующий:

John[*believes* →
  ${*Bob*[*likes* → ? *X*]: - *Mary*[*likes* → *?X*]}].

Это предложение материализует правило (не просто факт) и утверждает, что *John* также верит, что *Bob*'у нравятся все, кто нравится *Mary*. Соединяя с предыдущим оператором, что *John* верит в то, что *Mary* нравится *Sally*, можно было бы ожидать заключения, что *John* также поверит в то, что *Bob*'у нравится *Sally*. Однако такое заключение не может быть получено, поскольку неизвестно, является ли *John* способным к рассуждению существом, которое может употреблять modus ponens в повседневной жизни. Но эту информацию можно сообщить простым образом:

John[*believes* → *?A*]: -
  John[*believes* → {${*?Head*: - *?Body*}; *?Body*}].

FLORA-2 включает также транзакционную логику с некоторыми уточнениями, позволяющими отличать запросы от транзакций и обеспечить возможность реализации контроля в период компиляции. В транзакционной логике как действия, так и запросы представляются предикатами. В языке FLORA-2 транзакции представляются методами объектов с префиксом «%».

В случае если постусловие транзакции выполняется, изменения вносятся в базу данных. Если это условие не выполняется, текущая попытка изменения базы данных считается неприемлемой и предпринимается следующая попытка. Если не находится ни одного приемлемого выполнения, транзакция не выполняет никаких изменений в базе данных. Таким образом обеспечивается атомарность транзакций.

В языке FLORA-2 поддерживается концепция областей действия для отрицания по умолчанию. Модуль в языке FLORA-2 — это контейнер для отдельной базы знаний (или ее части). Модули позволяют изолировать отдельные части базы знаний и обеспечивают интерфейс для взаимодействия таких частей. Модули могут образовываться динамически, связываться с нужными частями базы знаний, обеспечивая эффективный вид инкапсуляции. Область действия запроса ограничивается модулем или всеми модулями, зарегистрированными в приложении. Это решение позволяет считать предположение о замкнутости мира (Closed World Assumption, CWA) приемлемым даже в среде Веба, если область действия программы явно и точно определена. Тем самым удается устранить возражения, направленные против идеи предположения о замкнутости мира, основанные на том, что Веб практически бесконечен и неудача в выводе некоторого факта из имеющейся информации не дает гарантии правильности заключения о том, что этот факт является ложным.

4 Примеры использования и требования к RIF

4.1 Примеры использования RIF и системы на правилах в области интересов группы RIF WG

При создании средств унификации неоднородных языков и обеспечения интероперабельности соответствующих систем [38] для уточнения требований к методам и средствам унификации и проверки их адекватности с самого начала проекта следует позаботиться об установлении взаимодействия с группами создания и поддержки конкретных разнородных языков и систем. Группой RIF WG изначально было определено, что система на правилах представляет интерес для RIF WG, если для нее будут взяты обязательства разработать применения, в которых потребовался бы обмен правилами с другими представляющими интерес для RIF WG системами на правилах, и реализовать соответствующий пример использования. Более 50 предложений примеров использования были получены RIF WG в 2005 г. [39].

Категоризация примеров использования, приведенная первоначально в [40], продолжает эволюционировать до сих пор. По мере развития проекта RIF требования к примерам использования уточнялись, в частности, для того, чтобы они в наибольшей мере соответствовали уже определенным диалектам. Результаты этой деятельности определены в уточняющих документах RIF WG [41].

Список категорий примеров использования, согласно [40], выглядит следующим образом.

1. Интеграция информации.
2. Принятие решений.
3. Кросс-платформенная разработка и развертывание правил.
4. Стратегии авторизации транзакций и управления доступом.
5. Обмен бизнес-правилами, ориентированными на взаимодействие с пользователем.
6. Публикация программ на правилах.
7. Сервисы обмена правилами для третьей стороны.
8. Развитое представление знаний.

Анализ полученных примеров использования показывает, что правила применяются для реализации разнообразных задач, и поэтому системы на правилах нельзя считать монолитными. Правила использовались для контроля качества вывода, реализации вычислений, управления информационными потоками, проверки ограничений целостности в базах данных, представления стратегий и управления ими, управления устройствами и процессами в реальном масштабе времени, определения потребности вмешательства человека в процессы управления и др.

4.2 Требования к RIF

RIF должен быть определен таким образом, чтобы обеспечивалась возможность создания новых диалектов (требование расширяемости) в соответствии с основными целями и общими требованиями RIF, равно как и возможность изменения существующих диалектов (при соблюдении требования совместимости снизу вверх).

Достижение междиалектной интероперабельности само по себе является плохо определенной задачей, поскольку известно, что трансляция диалектов с различной семантикой при полном сохранении смысла едва ли реализуема в большинстве случаев. Это не означает, что междиалектная трансляция вообще невозможна, однако требуются дополнительные критерии для формулировки точного понимания того, что считать удовлетворительной трансляцией (посредством обмена RIF-правилами). До сих пор определение критериев качества междиалектной трансляции не входило в задачи RIF WG.

Цели проекта и примеры использования предопределили требования к RIF. Требования, определенные как общие, предусматривают набор фундаментальных свойств, которыми должны обладать создаваемые диалекты.

4.2.1 Общие требования

1. *Реализуемость*. RIF должен быть реализуем на основе устоявшихся методов и не должен требовать дополнительных исследований, например алгоритмических или семантических проблем разработки трансляторов.
2. *Семантическая точность*. Ядро RIF должно иметь ясно и точно определенные синтаксис и семантику. Каждый стандартный диалект RIF должен иметь ясно и точно определенные синтаксис и семантику, расширяющие ядро RIF.
3. *Расширяемость формата*. Должна быть обеспечена возможность создания новых диалектов RIF, расширяющих существующие диалекты (совместимость в обратном направлении с уже созданными диалектами) и допускающих их постепенное внедрение в системах, которые используют уже существующие диалекты (совместимость в прямом направлении, в направлении развития).
4. *Трансляторы*. Для каждого стандартного диалекта RIF должна существовать возможность создания трансляторов между языками правил, определяемыми данным диалектом и RIF, без изменения языка правил.
5. *Стандартные компоненты*. Реализации RIF должны обеспечивать возможность использования стандартных вспомогательных технологий, таких как парсеры XML, генераторы парсеров, и не должны требовать специализированных реализаций при возможности повторного использования существующих решений.
6. *Охват спектра языков на правилах*. Из-за большого разнообразия языков на правилах едва ли найдется единственный язык обмена, который мог бы служить мостом для всех языков на правилах. Поэтому RIF предлагает диалекты, каждый из которых ориентирован на кластер подобных друг другу языков на правилах. RIF должен обеспечивать внутридиалектную интероперабельность, т. е. интероперабельность между семантически подобными языками на правилах (при помощи обмена правилами RIF) в рамках одного диалекта, а также он должен поддерживать междиалектную интероперабельность, т. е. интероперабельность между диалектами на максимально возможном их пересечении.

4.2.2 Требования, мотивированные примерами использования

1. *Модель совместимости*. Спецификации RIF должны определять четкие критерии конформности для идентификации реализаций RIF, которые являются конформными.
2. *Поведение по умолчанию*. RIF должен специфицировать на соответствующем уровне детализации поведение по умолчанию, которого можно ожидать от приложения, совместимого с RIF, но не обладающего способностью обработки всех или части правил, определенных в RIF-документе, либо он должен обеспечивать способ описания подобного поведения по умолчанию.
3. *Семантические различия*. RIF должен охватывать языки на правилах, имеющие различную семантику.

4. *Ограниченное число диалектов*. RIF должен иметь стандартное ядро и ограниченное число стандартных диалектов, базирующихся на этом ядре.
5. *Данные OWL*. RIF должен охватывать базы знаний на OWL как данные настолько, насколько они оказываются совместимыми с семантикой RIF.
6. *Данные RDF*. RIF должен охватывать триплеты RDF как данные настолько, насколько они оказываются совместимыми с семантикой RIF.
7. *Идентификация диалекта*. Семантика RIF-документа должна однозначно определяться его содержимым без привлечения данных извне.
8. *Синтаксис XML*. RIF должен поддерживать синтаксис XML в качестве основного нормативного синтаксического представления.
9. *Типы данных XML*. RIF должен поддерживать надлежащий набор скалярных типов данных с ассоциированными операциями в соответствии с их определениями в XML Schema part 2 и связанных с этим документом спецификациях.
10. *Слияние наборов правил*. RIF должен поддерживать возможность слияния наборов правил.
11. *Идентификация наборов правил*. RIF должен поддерживать возможность идентификации наборов правил.

Подробные описания примеров использования даны в [39–41].

5 Краткий обзор основных решений RIF

Согласно [7], RIF изначально был ориентирован на обмен правилами, а не на создание одного всеобъемлющего языка на правилах. В противоположность другим стандартам Семантического Веба, таким как RDF, OWL и SPARQL, с самого начала было ясно, что одним-единственным языком не удастся охватить все парадигмы на правилах, активно используемые для представления знаний и моделирования в сфере бизнеса. Было осознано, что даже обмен правилами сам по себе является сверхзадачей. Известные системы на правилах попадают в одну из трех широких категорий: системы на языках первого порядка, системы логического программирования, системы на активных правилах. Эти системы синтаксически и семантически имеют мало общего. Более того, даже в пределах одной парадигмы имеются большие различия между системами.

Такое разнообразие предопределило подход RIF WG, заключающийся в создании семейства языков, называемых диалектами, со строго определенными синтаксисом и семантикой. Семейство диалектов RIF должно быть однородным и расширяемым. Однородность означает, что диалекты должны впитать в себя как можно больше из существующего разнообразия синтаксического и семантического аппарата. Расширяемость означает, что при наличии обоснованной мотивации должна быть обеспечена возможность определения новых диалектов RIF как синтаксических расширений существующих диалектов, содержащих новые элементы с требуемой дополнительной функциональностью. Такие новые диалекты RIF, не являясь стандартами вначале, в дальнейшем могут быть стандартизованы.

Ввиду требования строгости определений слово «формат» в названии RIF звучит несколько уничижительно. Фактически RIF предоставляет больше, чем просто формат. Однако понятие формата является существенным для понимания способа предполагаемого использования RIF. В конечном счете в качестве среды обмена при помощи RIF между различными системами на правилах предлагается XML-формат обмена данными. Основная идея обмена правилами при помощи RIF заключается в том, что для различных систем будут определены отображения их собственных языков в диалекты RIF и обратно. Такие отображения должны сохранять семантику, так что полученные наборы правил могут передаваться от одной системы к другой при условии, что эти системы могут общаться друг с другом при помощи подходящего диалекта, поддерживаемого обеими системами.

5.1 Диалекты RIF

Работа RIF WG была сосредоточена на двух видах диалектов — диалектах, основанных на логике, и диалектах на правилах с действиями [7]. Вообще говоря, диалекты, основанные на логике, включают языки, использующие логику некоторого вида, такую как логика первого порядка (часто ограниченная логикой Хорна) или непервопорядковые логики, лежащие в основе различных языков логического программирования (например, логического программирования, следующего хорошо обоснованной или стабильной семантике). Диалекты на правилах с действиями включают системы на продукционных правилах, такие как Jess, Drools и JRules, и системы с реактивными правилами (или правилами «событие–условие–действие»), такие как Reaction RuleML и XChange. Из-за ограниченности ресурсов RIF WG были определены только два логических диалекта — базовый логи-

ческий диалект (RIF-BLD) и его подмножество: диалект-ядро RIF, расширяемое также диалектом продукционных правил (RIF-PRD). RIF-PRD — это единственный диалект на правилах с действиями, определенный RIF WG.

Диалект RIF-BLD соответствует логике Хорна с различными синтаксическими и семантическими расширениями[1]. Основные синтаксические расширения включают синтаксис фреймов и предикаты с поименованными аргументами. Основные семантические расширения включают типы данных и внешне определенные предикаты. Хотя этот диалект не является достаточно выразительным для многих применений правил, он охватывает большое число существующих систем на правилах, что позволяет считать его необходимой предпосылкой создания более выразительных диалектов в будущем. Эту деятельность предполагается вести в рамках каркаса расширения RIF, называемого RIF-FLD.

RIF-PRD — это другой разработанный RIF WG крупный диалект, охватывающий основные аспекты различных систем на продукционных правилах. Серьезный интерес к технологиям на продукционных правилах был проявлен крупными компаниями. Продукционные правила в соответствии с существующей практикой в системах, подобных Jess или JRules, определяются, используя неформальные вычислительные схемы, не основанные на логике. Поэтому RIF-PRD не является частью набора логических диалектов RIF. Вместе с тем значительное внимание было уделено тому, чтобы выделить максимально возможную общую часть продукционного и логических диалектов. Выделение такой общей части послужило основанием для создания диалекта-ядра RIF.

Предполагается, что существующие и будущие диалекты RIF будут использовать один и тот же набор типов данных, встроенных функций и предикатов в соответствии с их определением в документе RIF Datatypes and Built-Ins (RIF-DTB).

5.2 Каркас RIF для логических диалектов

Для упрощения создания новых диалектов и сокращения времени их разработки RIF WG определила каркас расширения RIF, называемый каркасом логических диалектов (RIF-FLD) [42]. Не исключено появление в будущем аналогичного каркаса правил с действиями. RIF-FLD не является собственно диалектом, он определен как универсальный каркас для создания новых логических диалектов, расширяющих существующие. Он был разработан для существенного упрощения процесса определения и верификации новых логических диалектов, расширяющих возможности RIF-BLD.

Разработка каркаса RIF-FLD оказалась возможной, поскольку, несмотря на различия логических теорий, составляющих основу различных логических систем на правилах, в них используется много общих синтаксических и семантических конструкций. Более того, способы образования комбинаций различных подобных конструкций для образования соответствующих систем хорошо изучены. Однако спецификация RIF-FLD уникальна, поскольку является результатом тщательного разбора этих знаний и представляет их в согласованном виде, причем XML используется даже на уровне каркаса.

RIF-FLD представляет собой весьма общий логический язык, использующий значительную часть широко используемых синтаксических и семантических конструкций: при этом, однако, намеренно ряд параметров оставлен неопределенным для того, чтобы конструкторы конкретных диалектов могли добавить необходимые детали. Например, RIF-FLD предоставляет способ представления правил синтаксиса при помощи понятия сигнатур. В нем также можно специфицировать некоторые семантические понятия, такие как модели и логическое следование, оставляя при этом открытым выбор конкретных вариантов (например, какие конкретно модели следует использовать при рассмотрении следования). Конструктор конкретного диалекта может определить его синтаксис и семантику заданием их специализации на основе синтаксиса и семантики RIF-FLD. При этом конструктор осуществляет возможность выбора параметров из вариантов, предоставляемых RIF-FLD, не требуя определения формул, типов данных, моделей, следования и пр. Такой подход продемонстрирован на определении RIF-BLD. Этот диалект специфицирован двумя способами: прямым перечислением

[1]Расширения языка трактуются следующим образом [25]. Пусть $L_1 \subseteq L_2$ — два логических языка, семантика которых определена на основе отношений следования (entailment) $\models_1$ и $\models_2$. Говорят, что L_2 является расширением L_1, если для любой пары формул φ, $\psi \in L_1$ следование $\varphi \models_1 \psi$ выполняется тогда и только тогда, когда выполняется $\varphi \models_2 \psi$. Для языков на правилах множество правил, которые можно использовать в роли посылок (premises), не совпадает с множеством правил, которые можно использовать в качестве следствий (consequents). Поэтому нужно предположить, что $L_1 = Premises_1 \cup Consequents_1$ и $L_2 = = Premises_2 \cup Consequents_2$. Более того, L_1 не обязательно должен быть подмножеством L_2. Скажем, он может включаться в L_2 на основе 1-1 трансформаций ι. В используемых обозначениях это можно выразить как $\iota(Premises_1) \subseteq Premises_2$ и $\iota(Consequents_1) \subseteq Consequents_2$. Теперь можно определить, что L_2 расширяет L_1 относительно трансформации ι, если для каждой пары формул $\varphi \in Premises_1$ и $\psi \in Consequents_1$ следование $\varphi \models_1 \psi$ выполняется тогда и только тогда, когда выполняется $\iota(\varphi) \models_2 \iota(\psi)$.

всех определений, что потребовало около 40 страниц текста, и заданием спецификации на основе RIF-FLD, что потребовало всего пяти страниц. Любое различие между двумя спецификациями следует рассматривать как ошибку, подлежащую исправлению. Таким образом, на примере двойной спецификации RIF-BLD показаны преимущества определения диалектов с помощью каркаса RIF-FLD.

Каркас RIF включает гибкий набор конструкций, достаточных для решения проблем:

- формирования диалектов с различной семантикой, включая хорошо обоснованную и стабильную семантику;
- введения правил, имеющих широкий диапазон представлений — от хорновских правил до универсальных правил с произвольными формулами в голове (таким образом, например, становятся представимыми как правила дизъюнктивного Datalog'a, так и GLAV (Global-and-Local-As-View) взгляды);
- выбора трактовки предположения об уникальности имен (в логическом программировании синтаксически разные базовые термы обозначают разные объекты, тогда как в дескриптивных логиках это не так);
- обеспечения возможности использования функций в правилах;
- использования развитых структур данных на основе фреймов;
- определения иерархии классов и отношений наследования;
- введения разнообразных скалярных типов данных (XML Schema).

Разработчики каркаса не считают его спецификацию незыблемой. Напротив, при необходимости в будущем она может быть изменена и расширена.

К моменту написания настоящей статьи было известно, что RIF-FLD был использован для определения трех логических диалектов: базового логического диалекта (RIF-BLD), диалекта-ядра средств программирования множества ответов (RIF-CASPD) и диалекта-ядра логического программирования, базирующегося на хорошо обоснованной семантике (RIF-CLPWD). Известно также, что RIF-FLD используется для определения диалекта правил с неопределенностью (Uncertainty Logic Dialect, ULD) и диалекта логики умолчания (Default Logic Dialect, DLD) как расширения BLD. Ведутся также работы по средствам синтаксического контроля текстов программ на соответствие различным диалектам, например на соответствие диалекту-ядру RIF.

5.3 Вопросы совместимости с RDF и OWL

Признавая необходимость сопряжения правил RIF с RDF и онтологиями OWL, RIF WG определила также концепции обеспечения совместимости RIF с RDF и OWL. Учитывая, что RIF, RDF и OWL — языки с различным синтаксисом и семантикой, следовало определить, как из правил RIF ссылаться на факты RDF и OWL, и выяснить, может ли иметь логический смысл всеобъемлющий язык. Специальный документ RIF, посвященный совместимости RIF с RDF и OWL, дает ответ на эти вопросы. Основная идея заключается в использовании синтаксиса фреймов для взаимодействия с RDF/OWL. Фреймы отображаются в триплеты RDF и определяется объединенная семантика для такой комбинированной конструкции. Обмен правилами посредством RIF может находиться в зависимости от данных RDF, RDF Schema или онтологий OWL либо использоваться вместе с ними. Таким образом, требуется определить возможность обеспечения интероперабельности RIF с другими стандартами Семантического Веба.

Гибридный подход к обеспечению интероперабельности правил и онтологий [25]. Требуется создание логической надстройки над стеками OWL и правил, которая позволяла бы использовать вывод, реализуемый в OWL в правилах, и, наоборот, вывод на основе правил использовать в OWL. Основная идея такого взаимодействия заключается в том, что системы на правилах и OWL будут рассматривать друг друга как черные ящики с интерфейсом, реализуемым на основе экспортируемых предикатов. Онтологии OWL будут экспортировать свои классы и свойства, а базы знаний на правилах будут экспортировать некоторые из определенных в них предикатов. Каждая из баз знаний будет иметь возможность ссылаться на предикаты в другой базе знаний, причем одной из возможных трактовок таких предикатов является их экстенсиональная интерпретация как множеств фактов. В частности, онтологии можно использовать как разделяемые многими приложениями концептуализации предметных областей. В этом случае различным приложениям в этой области могут соответствовать специфические программы на правилах.

Такой подход, называемый гибридным, обеспечивает возможность интегрированного использования существующих машин вывода в системах на правилах и в онтологических системах для реализации рассуждений в гибридном языке вместо того, чтобы создавать совершенно новую машину вывода [43].

Можно представить себе два языка — язык логического программирования R и онтологический

язык S (последний можно считать базированным на дескриптивной логике). Гибридное правило над R и S может иметь следующий вид:

$$H\text{: - } B_1, \ldots, B_m, Q_1, \ldots, Q_n.$$

Здесь $m, n \geq 0$; H, B_i — литералы, а Q_j — запросы, представленные на языке запросов Q_S.

Гибридная база знаний $K = (D, P)$ представляет собой конечное множество D аксиом в дескриптивной логике на онтологическом языке S и конечное множество гибридных правил P над R и S, включающих атомы не в дескриптивной логике. Оба языка поддерживаются машинами вывода, отвечающими на запросы на соответствующем языке запросов Q_R и Q_S. Язык запросов в гибридной парадигме синтаксически совпадает с языком запросов Q_R. Ответом на запрос является совокупность фактов, которая следует из гибридной базы знаний.

Однородный подход к обеспечению интероперабельности правил и онтологий [44]. В гибридном подходе обычные предикаты в языке на правилах и онтологические предикаты (которые чаще всего выступают в роли ограничений в посылках правил) строго разделены. Рассуждения осуществляются на основе сопряжения существующих машин вывода системы на правилах и онтологии.

В однородном подходе как онтологии, так и правила выражаются на одном языке L без необходимости установления априорных различий между предикатами на правилах и онтологическими предикатами.

5.4 Сценарии использования RIF совместно с данными RDF или онтологиями RDFS/OWL

Такие сценарии определены следующим образом [45]. Партнер по взаимодействию A использует язык правил, в котором имеются возможности для работы с RDF/OWL. Подобные возможности могут включать поддержку доступа к данным RDF, использование онтологий RDFS или OWL либо расширение RDF(S)/OWL средствами языка правил A. Используя RIF, A посылает партнеру B свои правила, возможно, содержащие ссылки на нужные графы RDF. Партнер B получает правила и извлекает требуемые графы RDF. Правила транслируются во внутренний язык B и обрабатываются вместе с RDF-графами, используя возможности для работы с RDF/OWL машины вывода B (подобные названным выше для партнера A).

Специализацией этого сценария является публикация правил RIF, ссылающихся на RDF-графы (публикация рассматривается как специальный вид взаимодействия «один ко многим»). Когда партнер A публикует свои правила в Вебе, может быть несколько потребителей, получающих RIF-правила и RDF-графы из Веба, транслирующих RIF-правила в соответствующие языки правил и обрабатывающих их вместе с RDF-графами в их собственных машинах вывода.

Другая возможная специализация сценария обмена правилами опирается на намерение расширения онтологии OWL правилами партнером, осуществляющим публикацию. Партнер по взаимодействию A использует язык правил, расширяющий OWL. Партнер A расщепляет свое определение онтологии и правил в раздельные онтологию OWL и документ RIF, публикует онтологию OWL и посылает (или публикует) документ RIF, который включает ссылки на онтологию OWL. Потребитель правил извлекает и транслирует онтологию OWL и документ в комбинированное описание, содержащее онтологию и правила в собственном языке правил, расширяющем OWL.

5.5 Состояние реализации RIF

RIF WG на основании отчетов по реализации RIF представила информацию в виде таблицы [46]:

Система на правилах	Организация	Диалект RIF	Поддержка функций
SILK	Vulcan, BBN, Stony Brook University	BLD, DLD	Producer, consumer
OntoBroker 5.3	Ontoprise	BLD (partial)	Producer, consumer
Fuxi	Chimezie Ogbuji	RIF Core and OWL 2 RL in RIF	Producer
IBM Websphere ILOG JRules	IBM/ILOG	PRD + Core	Producer (PRD), consumer (PRD + Core)
Eye	Jos De Roo	BLD + DTB	Consumer
VampirePrime	Alexandre Riazanov	BLD	Consumer
RIFle	José Maria Álvarez	Core, PRD, DTB	Validator
OBR	Oracle	PRD without import	Producer, consumer
IRIS	STI Innsbruck	BLD + DTB	Producer, consumer
N/A	Stijn Heymans, Michael Kifer	FLD	RIF-CASPD
N/A	Jidi Zhao, Harold Boley	FLD	RIF-URD
Riftr	Sandro Hawke	Core, BLD, DTB, RDF import	Producer, consumer, validator

6 Заключение

Работа по стандарту унификации языков на правилах, выполненная RIF WG, является значи-

тельным событием. Предложен тщательно обоснованный и формализованный подход к созданию семейства диалектов языков на правилах, позволяющий однородно представить многообразие существующих языков на правилах. Для логических языков на правилах определен каркас конструирования различных диалектов, позволяющих создавать взаимные сохраняющие семантику отображения различных логических языков существующих систем на правилах в такие диалекты. Такие отображения являются необходимой предпосылкой обеспечения интероперабельности и повторного использования логических программ. Новый диалект создается как расширение существующих диалектов, управляемое каркасом.

Впервые работа по унификации языков программирования выполняется как задача крупного консорциума. Само по себе осознание необходимости подобной работы таким разнородным сообществом, как W3C, уже является знаменательным событием.

Важным достижением является также выделение общего ядра для языков на правилах с совершенно различной парадигмой — логических и продукционных языков. Определение логического и продукционного диалектов как расширений ядра (первое из которых удовлетворяет каркасу) служит необходимой предпосылкой начала реализации RIF. Создание каркаса для совокупности продукционных языков — одна из задач развития RIF.

В результате выполненной RIF WG работы становится возможным рассмотрение программ на правилах как особого вида информационных ресурсов, которые могут быть использованы в различных средах интеграции и интероперабельного использования ресурсов (например, в средах поддержки предметных посредников). Благодаря высокому уровню абстракции и декларативности программ на правилах становится практически достижимой точная спецификация поведения предметных посредников (тогда как до сих пор при спецификации поведения на этом сугубо декларативном уровне приходилось ограничиваться спецификацией пред- и постусловий). В частности, применение различных моделей логических рассуждений становится реально возможным при решении задач с использованием множества неоднородных распределенных информационных ресурсов.

Литература

1. *Kalinichenko L. A.* Methods and tools for equivalent data model mapping construction // Advances in Database Technology: Conference (International) on Extending Database Technology EDBT'90 Proceedings, LNCS 416. — Berlin–Heidelberg: Springer-Verlag, 1990. P. 92–119.
2. *Kalinichenko L. A.* SYNTHESIS: The language for description, design and programming of the heterogeneous interoperable information resource environment. — 2nd ed. — Moscow: IPI RAN, 1993. 110 p.
3. *Kalinichenko L. A., Stupnikov S. A., Martynov D. O.* SYNTHESIS: A language for canonical information modeling and mediator definition for problem solving in heterogeneous information resource environments. — 4th ed. — Moscow: IPI RAN, 2007. 171 p.
4. *Захаров В. Н., Калиниченко Л. А., Соколов И. А., Ступников С. А.* Конструирование канонических информационных моделей для интегрированных информационных систем // Информатика и её применения, 2007. Т. 1. Вып. 2. С. 13–38.
5. *Abrial J. R.* The B-Book — assigning programs to meanings. — Cambridge: Cambridge University Press, 1996.
6. *Kifer M.* Rule interchange format: The framework // Web Reasoning and Rule Systems: 2nd Conference (International) Proceedings, LNCS 5348. — Berlin–Heidelberg: Springer Verlag, 2008. P. 1–11.
7. RIF Overview // W3C Working Draft. http://www.w3.org/TR/2010/WD-rif-overview-20100511.
8. List of rule systems. http://www.w3.org/2005/rules/wg/wiki/List_of_Rule_Systems.html.
9. RIF Working Group Web Page. http://www.w3.org/2005/rules/wiki/RIF_Working_Group.
10. *Reiter R.* A logic for default reasoning // Artificial Intelligence, 1980. Vol. 13. P. 81–132.
11. *Gallaire H., Minker J., Nicolas J.* Logic and databases: A deductive approach // ACM Computing Surveys, 1984. P. 153–185.
12. *Bidoit N., Froidevaux C.* Minimalism subsumes default logic and circumscription // LICS-87 Proceedings, 1987. P. 89–97.
13. *Gelfond M.* On stratified autoepistemic theories // AAAI-87 Proceedings, 1987. P. 207–211.
14. *Apt K. R., Blair H. A., Walker A.* Towards a theory of declarative knowledge // Foundations of deductive databases and logic programming. — Los Altos, CA: Morgan Kaufmann Publs., 1988. P. 89–148.
15. *Gelfond M., Lifschitz V.* The stable model semantics for logic programming // 5th Conference (International) on Logic Programming (ICLP) Proceedings. — 1988. P. 1070–1080.
16. *Besnard P.* An introduction to default logic. — Springer Verlag, 1989.
17. *Gelfond M., Lifschitz V.* Classical negation in logic programs and disjunctive databases // New Generation Computing, 1991. Vol. 9. P. 365–385.
18. *Marek W., Truszczyński M.* Autoepistemic logic // J. ACM, 1991. Vol. 38. No. 3. P. 587–618.
19. *Ramakrishnan R., Ullman J.* A survey of research on Deductive Database Systems: Technical report. — Stanford University, 1993.

20. *Minker J.* Logic and databases: A 20 year retrospective // LNCS 1154. — Springer-Verlag, 1996.
21. *Baral C.* Answer set programming: knowledge representation, reasoning and declarative problem solving using AnsProlog. — 2004. http://www.public.asu.edu/~cbaral/tutorial-dallas-june-04.pdf.
22. *Kowalski R.* Reasoning with conditionals in artificial intelligence. — Department of Computing, Imperial College London, 2009. http://www.doc.ic.ac.uk/~rak/papers/conditionals.pdf.
23. *Wan H., Grosof B., Kifer M., et al.* Logic programming with defaults and argumentation theories // Logic Programming, LNCS 564, 2009. P. 432–448.
24. Abduction. http://iph.ras.ru/page54852159.htm.
25. *Kifer M., de Bruijn J., Boley H., Fensel D.* A realistic architecture for the semantic web // 1st Conference (International) on Rules and Rule Markup Languages for the Semantic Web (RuleML2005) Proceedings, LNCS 3791. — Springer-Verlag, 2005. P. 17–29.
26. *Van Gelder A., Ross K. A., Schlipf J. S.* The well-founded semantics for general logic programs // J. ACM, 1991. Vol. 38. No. 3. P. 620–650.
27. *Pearce D.* A new logical characterization of stable models and answer sets // Non-monotonic extensions of logic programming: Lecture notes in artificial intelligence 1216. — Springer, 1997. P. 57–70.
28. *Eiter T., Ianni G., Krennwallner T.* Answer set programming: A primer // 5th International Reasoning Web Summer School 2009 Proceedings, LNCS 5689. — Springer-Verlag, 2009.
29. *Moore R. C.* Semantical considerations on nonmonotonic logic // Artificial Intelligence, 1985. Vol. 25. P. 75–94.
30. *Marek W., Truszczynski M.* Stable models and an alternative logic programming paradigm // The logic programming paradigm: 25 year perspective. — Springer Verlag, 1999. P. 375–398.
31. *Niemelä I., Simons P., Soininen T.* Stable model semantics of weight constraint rules // 5th Conference (International) on Logic Programming and Nonmonotonic Reasoning Proceedings. — Springer-Verlag, 1999.
32. *Lifschitz V.* What is answer set programming? // 23rd National Conference on Artificial Intelligence Proceedings. — AAAI Press, 2008. Vol. 3. P. 1594–1597.
33. System DLV. http://www.dbai.tuwien.ac.at/proj/dlv.
34. *Kifer M.* Nonmonotonic reasoning in FLORA-2 // Logic Programming and Nonmonotonic Reasoning, LNCS 3662. — Springer-Verlag, 2005. P. 1–12.
35. *Kifer M., Lausen G., Wu J.* Logical foundations of object-oriented and frame-based languages // J. ACM. — N.Y.: ACM New York, 1995. Vol. 42. Iss. 4. P. 741–843.
36. *Ludäscher B., Himmeröder R., Lausen G., May W., Schlepphorst C.* Managing semistructured data with Florid: A deductive object-oriented perspective // Information systems. — Elsevier, 1998. Vol. 23. Iss. 8. P. 589–613.
37. *Yang G., Kifer M., Wan H., Zhao C.* FLORA-2: User's manual, 2008. http://flora.sourceforge.net/docs/floraManual.pdf.
38. *Boley H., Kifer M., Patranjan P.-L., Polleres A.* Rule interchange on the Web reasoning // Web 2007, LNCS 4636. — Springer-Verlag, 2007. P. 269–309.
39. RIF Use Cases. http://www.w3.org/2005/rules/wg/wiki/Use_Cases.html.
40. General use case categories. http://www.w3.org/2005/rules/wg/wiki/General_Use_Case_Categories.
41. RIF Use Cases and Requirements // W3C Working Draft, 2008. http://www.w3.org/TR/rif-ucr.
42. RIF Framework for logic dialects // W3C Proposed recommendation / Eds. H. Boley, M. Kifer. — 2010. http://www.w3.org/TR/2010/PR-rif-fld-20100511.
43. *Grosof B. N., Horrocks I., Volz R., Decker S.* Description logic programs: Combining logic programs with description logic // WWW2003 Proceedings, 2003.
44. *Antoniou G., Damasio C. V., Grosof B., et al.* Combining rules and ontologies. A survey. Project REWERSE Report, 2005.
45. Implementations — RIF. http://www.w3.org/2005/rules/wiki/Implementations.
46. RIF RDF and OWL Compatibility // W3C Candidate recommendation. — 2009. http://www.w3.org/TR/2009/CR-rif-rdf-owl-20091001.

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2012. Т. 6. Вып. 2. С. 77–86

КОГНИТИВНЫЕ ИССЛЕДОВАНИЯ АССИСТИВНОГО МНОГОМОДАЛЬНОГО ИНТЕРФЕЙСА ДЛЯ БЕСКОНТАКТНОГО ЧЕЛОВЕКО-МАШИННОГО ВЗАИМОДЕЙСТВИЯ*

А. А. Карпов[1]

Аннотация: Представлены результаты исследований многомодального пользовательского интерфейса, предназначенного для бесконтактного управления персональным компьютером при помощи речевого ввода и указательных жестов/движений головой. Данный многомодальный интерфейс использует низкобюджетное аудио- и видеооборудование для одновременного захвата многоканальных сигналов и обеспечивает универсальный доступ к компьютерным системам как обычных операторов для бесконтактной (без использования рук) работы с компьютером, так и пользователей с ограниченными физическими возможностями (с проблемами двигательных функций рук или даже не имеющих рук/пальцев). Описаны методики и результаты количественной оценки производительности бесконтактного человеко-машинного взаимодействия с применением элементов когнитивных экспериментов и сравнение с результатами для стандартных контактных способов указательного ввода информации.

Ключевые слова: многомодальный интерфейс; распознавание речи; машинное зрение; ассистивные информационные технологии

1 Введение

Многие люди не могут полноценно работать с компьютерными системами (печатать тексты, работать в Интернете, рисовать, и т. д.) по причине физических ограничений, например ампутации рук в результате войн, аварий, врожденных дефектов или парализации рук в результате болезней. Для таких людей и создается многомодальный пользовательский интерфейс бесконтактного взаимодействия с компьютером посредством речевого ввода и отслеживания осмысленных движений (жестов) головы или тела человека. Согласно общепринятому определению, «жест» (от лат. *gestus* — движение тела) — это некоторое действие или движение человеческого тела или его части (например, рук, головы или глаз), имеющее определенное значение или смысл. В этом смысле жестом может являться кивок, покачивание или наклон головы, а также указательные жесты, когда пользователь головой показывает на определенное направление движения.

Разработанный за последние годы в лаборатории речевых и многомодальных интерфейсов СПИИРАН ассистивный (предназначенный для помощи) пользовательский интерфейс получил сокращенное название ICanDo («Я могу делать»), что расшифровывается как “Intellectual Computer AssistaNt for Disabled Operators” («Интеллектуальный компьютерный помощник для операторов-инвалидов») [1]. Он снабжен программными технологиями автоматического распознавания русской речи / голосовых команд и технического зрения для отслеживания движений головы (указательных жестов) с целью управления курсором мыши на экране дисплея, что повышает естественность и эффективность человеко-машинного взаимодействия. Речевое взаимодействие является наилучшей альтернативой любым устройствам ввода для задачи набора текста на компьютере как для пользователей-инвалидов, так и для обычных пользователей. Видео- и аудиосигналы одновременно и параллельно захватываются одним аппаратным устройством — цифровой видеокамерой (веб-камерой) и синхронно обрабатываются в многомодальном интерфейсе.

Альтернативой программному человеко-машинному интерфейсу для пользователей без верхних конечностей могут служить различные аппаратные устройства для управления графическим интерфейсом компьютера, например аппаратно-программные устройства слежения — трекеры головы (зарубежная система InterTrax, которая использует гироскоп; система SmartNav, которой необходим инфракрасный приемопередатчик;

* Работа выполнена при поддержке Минобрнауки РФ в рамках ФЦП «Исследования и разработки», госконтракт № 11.519.11.4025; Совета по грантам Президента РФ, проект МК-1880.2012.8; фонда «Научный Потенциал» и Комитета по науке и высшей школе Правительства Санкт-Петербурга.

[1] Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН), karpov@iias.spb.su

оптическая система HeadMouse Extreme). Чтобы использовать данные системы для управления курсором мыши на экране дисплея, пользователь должен надеть на голову специальное устройство (шлем или очки виртуальной реальности со встроенным микроминиатюрным гироскопом в случае InterTrax, либо специальную конструкцию со светоотражающими метками в случае SmartNav или HeadMouse Extreme). Кроме того, для этой задачи могут также применяться специальные устройства со светодиодами и аккумуляторами, например комплект для ассистивного управления компьютером КАУ-09-1, разработанный в ЗАО НПК ФАТУМ, или цветными реперными (контрольными) точками-мишенями, которые крепятся на специальном шлеме, надеваемом на голову, например аппаратная система «Шлемомышь» [2], разработанная лабораторией компьютерной графики факультета вычислительной математики и кибернетики МГУ. Реперные точки на таких устройствах отслеживаются посредством инфракрасной либо цифровой видеокамеры. Однако пользователи и психологи говорят о том, что люди не хотят использовать для человеко-машинного взаимодействия специальные, носимые на голове или теле аппаратные устройства, значительно снижающие естественность взаимодействия и мобильность передвижения из-за наличия проводов, кабелей, аккумуляторов для автономной работы, их общей громоздкости и технических сложностей в калибровке и установке. Кроме того, люди без рук не могут надеть такое устройство сами себе на голову.

Для некоторых задач и для определенных категорий пользователей-инвалидов (например, парализованных лежачих людей) перспективно применение аппаратно-программных систем для трекинга направления взгляда. Такие системы в мире существуют, например зарубежные трекеры глаз Eyegaze System или Visual Mouse, но их использование и внедрение на практике осложняется тем, что необходимо использовать очень дорогие высокоскоростные цифровые видеокамеры высокой четкости с большим разрешением, так как область глаза незначительна по размеру и сложна в распознавании. Кроме того, как показывают когнитивные исследования [3], использование отслеживания направления взгляда для управления курсором мыши намного сложнее для обучения и хуже, чем отслеживание движений головы, по следующим показателям: производительность, эмоциональная нагрузка на пользователя, удобство использования, эргономичность.

Кроме того, в качестве альтернатив бесконтактному взаимодействию можно упомянуть управление манипулятором-мышью с использованием ног вместо рук или специальный тактильный манипулятор, функционирующий за счет изменения положения центра масс тела человека, сидящего на специальной «подушке» [4]. В будущем, возможно, будут доступны и системы взаимодействия на основе прямого интерфейса мозг–компьютер, во всяком случае, разработки в области нейроинформатики активно ведутся как за рубежом, так и в России.

В ассистивном интерфейсе ICanDo, которому посвящена данная статья, реализованы и применены программные средства компьютерного зрения для обнаружения лица человека в оптическом потоке на основе характерных органов/черт лица (нос, глаза, губы) без использования искусственных маркеров/мишеней и специализированных, носимых человеком, устройств, что выгодно отличает его от имеющихся аналогов. Программная система взаимодействия не накладывает дополнительных ограничений на пользователя и обеспечивает естественность и комфорт при бесконтактной работе с компьютером. Применяемые в интерфейсе голосовые команды, распознаваемые автоматической системой, являются прекрасной альтернативой стандартным органам ввода информации (клавиатура) как для инвалидов без рук или пальцев рук, так и для обычных пользователей.

2 Архитектура ассистивного многомодального интерфейса пользователя

Ассистивный интерфейс человеко-машинного взаимодействия относится к классу многомодальных пользовательский интерфейсов [5] и использует две естественные входные модальности: речь на русском языке и указательные жесты — движения головы (вверх, вниз, вправо, влево и в любых промежуточных направлениях). Обе модальности являются активными [6] и инициируются напрямую человеком, поэтому они непрерывно отслеживаются и обрабатываются интеллектуальными подсистемами интерфейса. Каждая из модальностей передает свою семантическую информацию: положение головы определяет положение курсора мыши на рабочем столе компьютера в конкретный момент времени, а речевой сигнал передает информацию о действии, которое должно быть выполнено с некоторым объектом графического пользовательского интерфейса. На рис. 1 представлена архитектура аппаратно-программного комплекса ассистивного многомодального интерфейса.

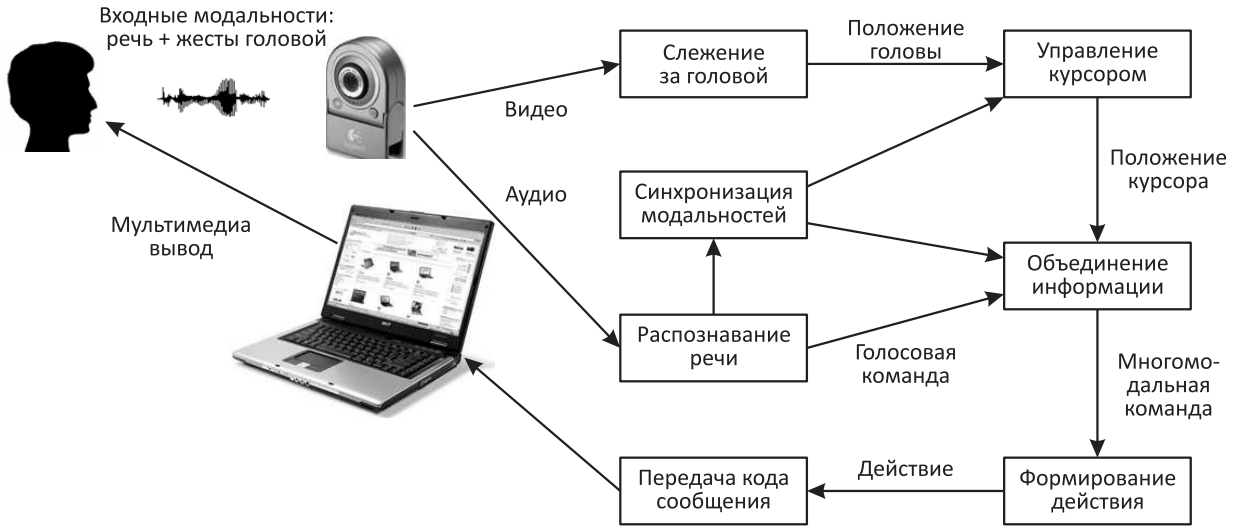


Рис. 1 Архитектура ассистивного многомодального пользовательского интерфейса

Многомодальный интерфейс способен распознавать несколько десятков голосовых команд для управления компьютером (например, «открыть», «сохранить», «левая», «правая», «ввод» и т. д.). Всего система содержит 40 голосовых команд, которые являются наиболее часто используемыми командами при работе с графическим пользовательским интерфейсом русскоязычного варианта операционной системы семейства Microsoft Windows. Теоретически возможно работать с компьютером, используя лишь левую и правую кнопку мыши (команды «левая» и «правая»), однако введение дополнительных голосовых команд позволяет серьезно ускорить и упростить процесс человеко-машинного взаимодействия и производительность работы.

Подаваемые голосовые команды захватываются встроенным в видеокамеру дистанционным микрофоном, передаются в цифровом виде в компьютер и интерпретируются автоматической системой распознавания русской речи. Помимо этой системы для управления курсором мыши используется технология компьютерного зрения, отслеживающая перемещения головы пользователя. Координаты курсора мыши «привязываются» к координатам кончика носа и иных лицевых органов, автоматически определяемых на изображении, таким образом любые движения головы вызывают смещение курсора на экране в соответствующем направлении. Объединение такого интерфейса с речевым интерфейсом позволяет пользователям не только бесконтактно работать с графическим интерфейсом компьютера, подавая отдельные команды голосом, но и набирать текст в любом существующем редакторе или форме, используя виртуальную клавиатуру или произнося текст по буквам. Альтернативно для ввода текста может использоваться интерфейс Dasher [7], который является сторонней ассистивной программой для указательного ввода букв на разных языках.

2.1 Автоматическая обработка аудиовизуальных сигналов

Интерфейс способен обрабатывать одноканальный аудиосигнал от микрофона и распознавать голосовые команды на русском языке, разработаны также аналогичные версии для английского и французского языка. Для распознавания русской речи применяется оригинальная система автоматического распознавания речи, получившая название SIRIUS (SPIIRAS Interface for Recognition and Integral Understanding of Speech) [8]. В системе для параметризации звука используется разновидность спектральной обработки сигнала — мел-частотные кепстральные коэффициенты с их первой и второй производными. Акустическое моделирование звуков речи в системе производится с применением непрерывных скрытых марковских моделей (СММ) первого порядка [9] и смесей нормальных (гауссовских) распределений плотностей вероятностей векторов наблюдений в состояниях СММ. Для лучшего учета вариативности разговорной речи каждое слово преобразуется в последовательность произносимых фонем (звуков речи) и строится вероятностная модель для каждой фонемы. С помощью алгоритма Витерби вычисляется вероятность принадлежности последовательности векторов наблюдений СММ некоторого слова [9].

Для задачи голосового управления (работа с персональным компьютером относится к этой катего-

рии приложений), где применяется малый словарь распознавания, лексикон системы представляет собой линейный список всех команд с их фонематическими транскрипциями и может достаточно просто дополняться. Все голосовые команды ICanDo можно условно разделить на четыре класса по их функциональному назначению:

(1) команды, заменяющие управление кнопками и регуляторами манипулятора-мыши (например, «левая», «правая», «двойной клик», «прокрутка вниз» и т. д.);

(2) команды, заменяющие нажатие клавиш клавиатуры (например, «ввод», «удалить», «регистр», цифры, буквы и т. д.);

(3) команды управления графическим пользовательским интерфейсом (например, «открыть», «сохранить», «печать», «пуск» и т. д.);

(4) специальные команды («калибровка»).

Нужно отметить, что лишь команды, заменяющие работу мыши, являются фактически многомодальными, так как они используют информацию о положении курсора мыши в текущий момент времени. Все остальные являются исключительно речевыми одномодальными командами, и при их выполнении положение курсора не учитывается.

В многомодальном интерфейсе для управления курсором мыши используется подсистема компьютерного зрения, отслеживающая указательные движения головы пользователя. Применяется программный модуль для отслеживания движений головы пользователя, реализованный на основе базового алгоритма Лукаса–Канаде (Lukas–Kanade) [10] и его более поздней пирамидальной модификации [11] для анализа оптического потока, т. е. изображение видимого движения объектов, поверхностей или краев сцены, получаемое в результате перемещения наблюдателя относительно сцены или, наоборот, сцены относительно наблюдателя. В системе производится автоматическое отслеживание пяти естественных точек на лице: центр верхней губы, кончик носа, точка между глаз, зрачок правого глаза и зрачок левого глаза. Первоначальный поиск головы человека на статических изображениях (последовательных видеокадрах с разрешением 640×480 пикселей и частотой до 25 кадров в секунду, получаемых от видеокамеры) производится методом AdaBoost с применением алгоритма Виолы–Джонса (Viola–Jones) [12]. Изображение сканируется рамкой-окном заданного размера и строится пирамида копий объектов. Построенная пирамида анализируется заранее обученными каскадами Хаара, и на изображении находятся графические области, отвечающие заданной визуальной модели [13]. Реализованный метод детекции головы находит прямоугольные графические области на изображении, с высокой степенью вероятности содержащие изображение лица человека. Введено ограничение: размер такой области должен быть не менее 220×250 пикселей, чтобы захватывать только одно лицо в кадре, достаточно близко расположенное по отношению к видеокамере, а кроме того, это ускоряет процесс обработки видеопотока.

В отличие от имеющей аналогичное предназначение канадской системы Nouse [14], в которой отслеживается только положение кончика носа для управления движением курсора мыши, в ICanDo для более робастного слежения за перемещением головы оператора используется набор из 5 естественных лицевых объектов.

2.2 Синхронизация сигналов и объединение информации

В интерфейсе для объединения информации и выполнения многомодальной команды необходимо учитывать координаты указателя мыши, актуальные для момента времени непосредственно перед произнесением голосовой команды пользователем, т. е. должна сохраняться определенная история координат положения курсора. Если же использовать координаты указателя, актуальные на момент окончания произнесения голосовой команды, то многомодальная команда может оказаться неверной, так как курсор может сместиться от запланированного положения из-за непроизвольных перемещений головы (а они всегда существуют при говорении). В этом аспекте состоит принципиальное отличие указателя, управляемого движениями головы, от управляемого аппаратными манипуляторами наподобие мыши, трекбола, сенсорного экрана и т. д.

Звуковой сигнал, непрерывно захватываемый дистанционным стационарным микрофоном и передаваемый в компьютер посредством звуковой платы, обрабатывается модулем автоматического распознавания речи. Процесс распознавания речи запускается программным модулем детекции границ речи, который обнаруживает наличие в звуковом сигнале речевого фрагмента, отличного от тишины или постоянного фонового шума. Процесс распознавания заканчивается после получения наилучшей гипотезы распознавания голосовой команды из автоматической системы. Синхронизация модальностей производится следующим образом: текущее положение курсора сохраняется в буфере системы в первый момент определения наличия речи оператора (срабатывания алгоритма поиска границ речи по значению энергии сегментов

сигнала). По окончании процесса распознавания команды модуль распознавания речи дает сигнал на объединение информации и выполнение многомодальной команды. Таким образом, именно модуль распознавания речи осуществляет синхронизацию модальностей в бимодальном интерфейсе.

Для объединения информации, поступающей от двух модальностей, используется фреймовый метод позднего объединения, когда поля определенной структуры (фрейма) заполняются данными по мере их поступления, а по окончании процесса распознавания выполняется многомодальная команда. Поля семантического фрейма следующие: текст голосовой команды, абсцисса точки положения указателя мыши, ордината точки положения указателя, тип речевой команды (многомодальная или одномодальная). Если распознанная команда является многомодальной, она объединяется в единую команду с сохраненными координатами курсора и автоматически отсылается сообщение виртуальному устройству мыши о выполнении нужного действия. Если же голосовая команда одномодальна, то посылается соответствующее сообщение виртуальному устройству клавиатуры с кодом клавиши или сочетанием кодов. Движения головы сами по себе не могут подавать команд управления графическим пользовательским интерфейсом, однако они могут использоваться, например, для создания изображений в графических редакторах.

3 Когнитивные исследования пользовательского интерфейса

При помощи многомодального интерфейса ICanDo был проведен ряд экспериментов, которые были ориентированы на изучение организации бесконтактного взаимодействия человека с машиной и использовали элементы когнитивных исследований.

3.1 Методика исследований

Экспериментально была проведена оценка скорости и производительности работы пользователей с бесконтактным интерфейсом при указании на объекты графического пользовательского интерфейса. Для оценки скорости ввода информации была использована методология международного стандарта ISO 9241-9:2000 “Requirements for non-keyboard input devices” (Требования к неклавитаурным устройствам ввода информации) [15], которая базируется на экспериментах и законах, разработанных в середине XX в. американским психологом-когнитивистом Полом Фиттсом (Paul Morris Fitts) и впоследствии развитых другими учеными [16]. Применяемая в данном исследовании методика оценки интерфейса состоит в следующем. Тестеры, используя предоставленное им устройство указательного ввода, должны насколько возможно быстро отмечать на экране набор целей-объектов (последовательно кликнуть на них, т. е. дать голосовую команду «левая» для нажатия левой кнопки мыши), появляющихся по круговой схеме на мониторе. При этом порядок целей задается программой автоматически таким образом, чтобы пользователь последовательно выделял наиболее удаленно расположенные друг от друга объекты, совершая движения указателем в различных направлениях [17]. Когда нажатием на кнопку происходит подтверждение выделения текущего объекта-цели на экране, отображается следующая цель. При этом автоматически вычисляется индекс сложности задачи ID (*index of difficulty*), измеряемый в битах согласно формуле [18]:

$$\mathrm{ID} = \log_2\left(\frac{D}{W} + 1\right), \qquad (1)$$

где D — расстояние между центрами целей; W — диаметр цели.

Однако координаты точки, где происходит щелчок кнопкой мыши, зависят как от фактического (*effective*) расстояния между точками кликов, так и от фактического диаметра самих целей (т. е. чем меньше цель, тем сложнее попасть по ее центру). Поэтому фактический индекс сложности выражается следующей формулой [18]:

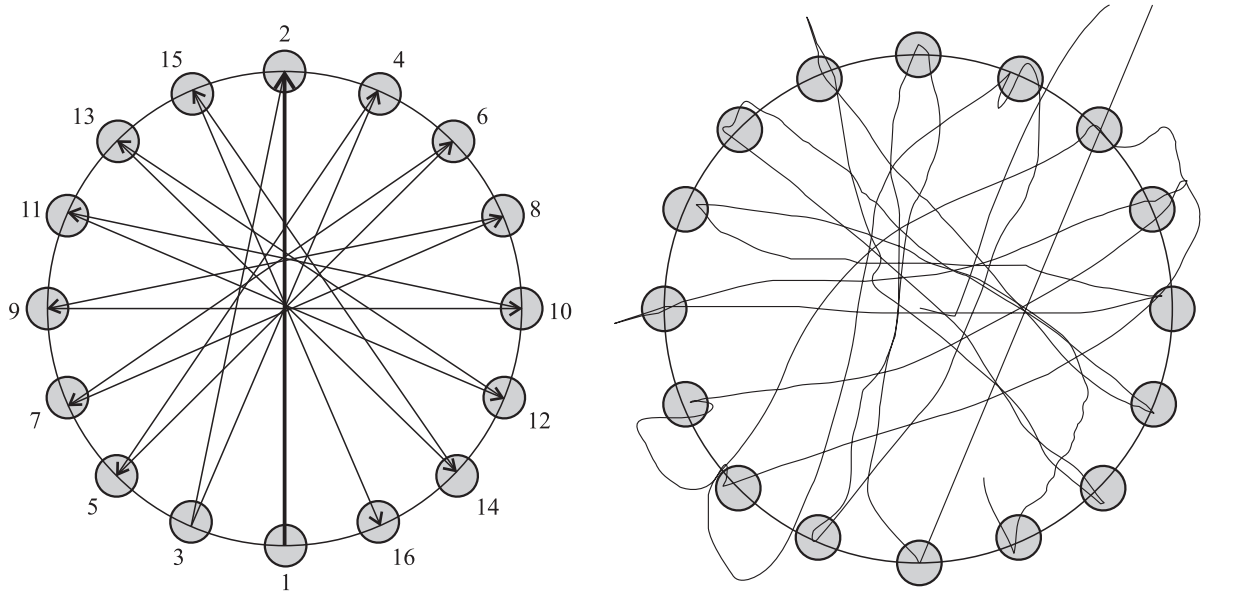
$$\mathrm{ID}_e = \log_2\left(\frac{D_e}{W_e} + 1\right). \qquad (2)$$

Здесь D_e — фактическое расстояние между точками кликов двух последних целей; W_e — фактический диаметр (или ширина) цели, определяемый в [18] как

$$W_e = 4{,}133\sigma, \qquad (3)$$

где σ — среднеквадратическое отклонение координат точки выделения (клика), проецируемой на ось, которая соединяет центры начальной и конечной целей. Получаемые значения ID_e отличаются от значений ID, более точно учитывая качество выполнения тестового задания пользователем.

Для проведения эксперимента было разработано соответствующее программное обеспечение, которое позволяет произвольно задавать значения D и W, а также фиксировать результаты прохождения теста. Программа для ЭВМ предлагает пользователю последовательно кликнуть на 16 целей, которые по очереди появляются на экране согласно



(*а*) (*б*)

Рис. 2 Схема и порядок расположения целей на экране для проведения когнитивных экспериментов с интерфейсом по методу Фиттса (*а*) и реальный пример траектории движения курсора мыши на экране при бесконтактном выполнении задания (*б*)

рис. 2, *а*. На рис. 2, *б* показан реальный пример получившейся траектории движения курсора мыши на экране при бесконтактном выполнении задания посредством ICanDo. Здесь можно видеть, что данная задача для пользователя не была простой, но ошибок выделения (непопаданий по целям) он не допустил.

3.2 Анализ результатов экспериментов

Для выполнения тестового задания были привлечены четыре пользователя-новичка, не имевших ранее опыта работы с многомодальным интерфейсом, и два пользователя-эксперта, принимавших участие в ее разработке и отладке. Каждым пользователем были проведены серии по 10 тестов с последовательным изменением диаметра цели W в пределах от 32 до 128 пикселей и среднего расстояния D между целями в пределах 96–650 пикселей (использовалось разрешение экрана 1280×1024), т. е. показатель ID варьировался от 1,32 до 4,4 бит. Каждый тест занимал в среднем 30–60 с.

Рисунок 3 показывает полученный в результате экспериментов и усредненный по всем пользователям график зависимости отношения значений ID_e (фактический индекс сложности) и ID (теоретически рассчитанный индекс сложности) при разных значениях D и W. Характерно, что данный график лежит выше пунктирной линии-нормали (ожидаемый теоретически индекс сложности выполнения задачи), а это означает, что выполнение данной задачи оказалось несколько сложнее, чем планировалось. В противном случае, если бы график зависимости лежал ниже нормали, то можно было бы говорить о том, что предлагаемая тестерам задача оказалась легче расчетной сложности.

Согласно экспериментам по методике Фиттса, время движения MT (*movement time*) между двумя целями линейно зависит от индекса сложности ID [19]. Полученное в ходе экспериментов среднее значение MT для всех тестеров равнялось 2550 мс, т. е. около 2,5 с между речевыми «нажатиями» цели. Рисунок 4 показывает два аппроксимирующих графика зависимости времени движения MT от фактической сложности задачи ID_e отдельно для пользователей-новичков, не работавших ранее с интерфейсом, и для обученных пользователей-экспертов. Хорошо заметно, что эффект обучения положительно сказывается на увеличении скорости бесконтактной работы с компьютером. Также разброс значений MT для новичков оказался значительнее, они выполняли тесты менее стабильно. На основании результатов экспериментов можно сказать, что новички начинают уверенно работать с компьютером бесконтактно при помощи многомодального интерфейса уже через 10–15 мин тренировки (исключая этап настройки системы распознавания речи на голосовые характеристики

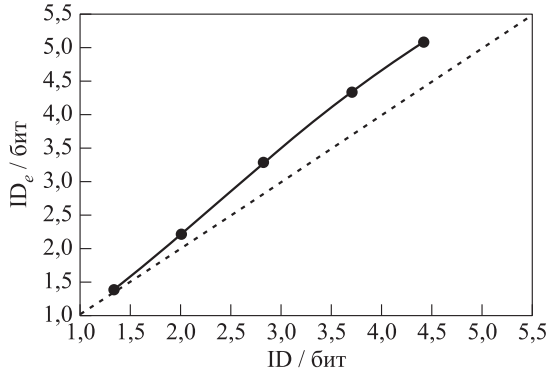


Рис. 3 График зависимости значений фактической сложности ID_e и теоретической сложности ID выполнения задачи и его отклонение от нормали

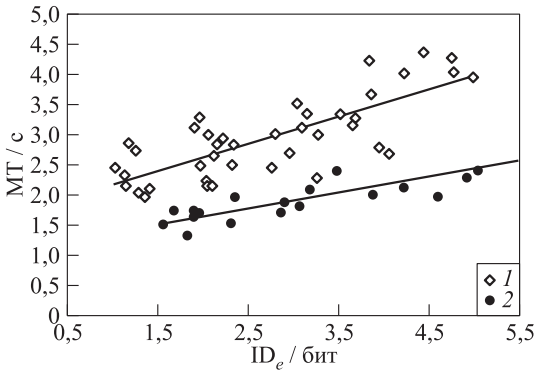


Рис. 4 Графики зависимостей времени движения MT от фактического индекса сложности ID_e задачи отдельно для новичков (*1*) и экспертов (*2*)

пользователя), что, конечно же, несколько больше, чем при первоначальном овладении мышкой и клавиатурой. Однако через день работы с системой пользователь уже может считаться экспертом в бесконтактном человеко-машинном взаимодействии.

В применяемой методике экспериментов Фиттса основным показателем оценки интерфейса является общая производительность работы пользователя с системой TP (*throughput*) [20], определяющая компромисс между временем движения (скоростью выполнения задания) и точностью выделения цели и измеряемая в битах в секунду согласно следующей формуле:

$$\mathrm{TP} = \frac{\mathrm{ID}_e}{\mathrm{MT}}. \qquad (4)$$

Полученное в ходе экспериментов среднее значение TP для всех тестеров составило 1,2 бит/с, максимальное значение TP для одного тестера — 2,0 бит/с.

Также в ходе когнитивных исследований была проведена сравнительная оценка контактных устройств для ввода/указания, таких как сенсорный экран 17″, джойстик, трекбол, сенсорная панель (*touchpad*) 3″ и стандартный манипулятор-мышь. Двумя пользователями были проведены серии по 10 тестов для каждого устройства с последовательным изменением диаметра цели W в пределах от 32 до 128 пикселей и среднего расстояния D между целями в пределах от 96 до 650 пикселей. Таблица 1 приводит результаты экспериментов и сравнения всех вышеуказанных устройств по трем основным количественным критериям:

(1) среднее время движения MT между двумя целями;

(2) процент ошибок выделения целей (непопадание курсором в цель);

(3) общая производительность указательного интерфейса TP.

Таблица 1 показывает, что наилучшие результаты по производительности интерфейсов были показаны сенсорным монитором, так как рука тестера свободно перемещается по воздуху. Управление курсором посредством многомодального интерфейса, отслеживающего движения головы, уступает по производительности практически всем аппаратным контактным средствам ввода информации, кроме джойстика (который весьма непривычен для управления курсором), однако имеет то преимущество, что является бесконтактным способом управления курсором и может применяться категориями потенциальных пользователей, для которых стандартные средства ввода информации недоступны.

Таблица 1 Сравнительная оценка эффективности интерфейсов для указательного ввода информации с использованием методики Фиттса

Устройство ввода	MT, с	Ошибка выделения, %	TP, бит/с
Джойстик	2,01	7,00	1,54
Трекбол	1,03	3,83	3,51
Сенсорная панель 3″	0,85	4,50	3,72
Манипулятор-мышь	0,49	3,17	6,65
Сенсорный экран 17″	0,50	6,17	7,85
Интерфейс ICanDo	1,98	7,33	1,59

Тестирование интерфейса в реальной задаче бесконтактной работы с компьютером было также проведено тремя добровольными пользователями. Пользователям предлагался определенный сценарий — последовательность операций, которую

Таблица 2 Сравнение бесконтактного и контактного интерфейсов человеко-машинного взаимодействия

Точность распознавания голосовых команд, %	Время выполнения тестового сценария, с	
	Интерфейс ICanDo	Мышь + клавиатура
96	82	43

пользователи должны были выполнить двумя способами (многомодальным — посредством ICanDo и стандартным — при помощи манипулятора-мыши). Тестовая задача включала в себя элементарные операции с текстовым редактором MS Word, а также поиск заданной информации в Интернете посредством MS Internet Explorer. Конкретнее: пользователю нужно было найти информацию о программе передач на интернет-портале Рамблер, скопировать интересующий фрагмент этой страницы, открыть текстовый редактор MS Word, вставить в пустой документ информацию из буфера, сохранить файл на рабочем столе и распечатать данный файл. Таблица 2 показывает количественные результаты экспериментов и сравнение двух способов человеко-машинного взаимодействия (среднее время, требуемое для выполнения всего тестового сценария и точность распознавания речи в дикторозависимом режиме).

Многомодальный бесконтактный способ ввода оказался в 1,9 раз медленнее, чем стандартный контактный способ, что было очевидно. При этом точность распознавания голосовых команд составила свыше 96% в дикторозависимом режиме работы. Однако, если учесть, что аудиосигнал, получаемый от встроенного в видеокамеру микрофона, характеризуется невысоким отношением сигнал/шум (SNR, signal-to-noise ratio), то полученный результат по точности распознавания можно считать приемлемым. Полученная скорость работы бесконтактного интерфейса вполне достаточна, так как он разрабатывается для помощи людям с физическими ограничениями, в частности для людей без рук или с парализованными руками.

Также был проведен анализ статистики бесконтактной работы пользователя-эксперта с интерфейсом ICanDo в течение одного дня в задаче навигации (серфинга) в Интернете посредством браузера MS Internet Explorer. Последующий анализ журнала статистики показал, что всего пользователь сделал более 750 голосовых команд, при этом некоторые команды были более частотными, чем другие, а некоторые команды не использовались вовсе. Диаграмма на рис. 5 показывает распределение частотности голосовых команд, примененных пользователем.

Легко было предсказать заранее, что наиболее популярной окажется команда «левая» (клик левой кнопкой мыши), которая использовалась более чем в трети случаев, включая и ввод текста при помощи специального программного обеспечения — экранной виртуальной клавиатуры. Однако необходимо сказать, что при работе с мышкой и клавиатурой это значение еще выше для подобной задачи, так как, работая бесконтактно, пользователи стараются избегать работы со сложными многоуровневыми меню стандартных офисных прикладных программ, заменяя их «горячими клавишами» для быстрого доступа к действиям. Все остальные команды рас-

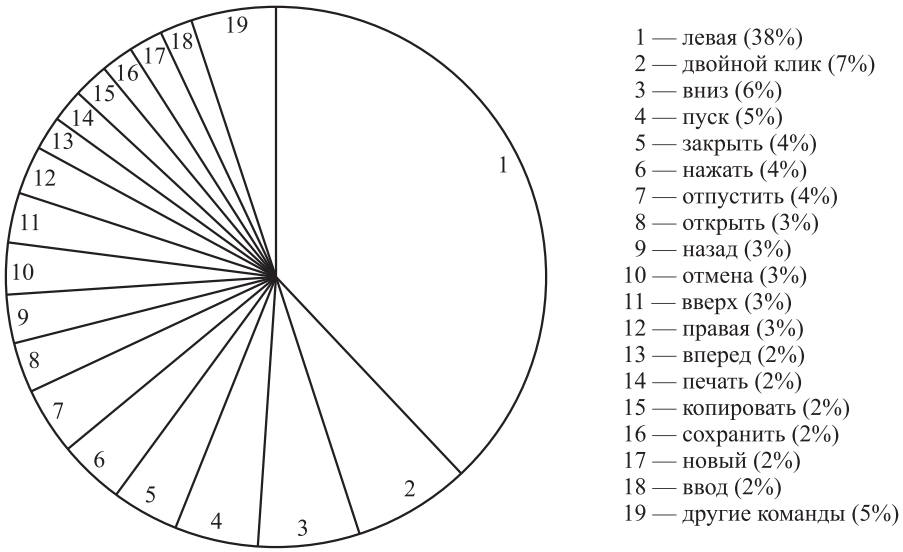


Рис. 5 Распределение относительной частоты использования голосовых команд тестером в ходе эксперимента

пределены более-менее равномерно среди оставшихся 62%. При этом 64% всех голосовых команд было подано многомодально (совместно с движениями головы для выделения графических объектов или ссылок на экране), а оставшиеся 36% команд — одномодально.

Видеодемонстрации бесконтактной работы пользователей с компьютером, в том числе и одного человека, не имеющего верхних конечностей, посредством ассистивного многомодального интерфейса ICanDo можно посмотреть на интернет-сайте лаборатории речевых и многомодальных интерфейсов СПИИРАН [21].

4 Заключение

В статье представлены результаты исследований бесконтактного человеко-машинного взаимодействия, реализуемого посредством ассистивного многомодального интерфейса ICanDo, предназначенного специально для работы человека-оператора с ЭВМ без использования рук. Описана общая архитектура ассистивного многомодального интерфейса, автоматическая обработка аудио- и видеосигналов, а также механизмы синхронизации и объединения модальностей. В данном ассистивном пользовательском интерфейсе для робастного отслеживания указательных жестов/движений головы оператора используется массив из пяти естественных точек на лице: центр верхней губы, кончик носа, точка между глаз, зрачок правого глаза и зрачок левого глаза. Применяются голосовые команды для бесконтактного управления прикладным и системным программным обеспечением компьютера. Результаты проведенных исследований с использованием методики Фиттса и элементов иных когнитивных экспериментов позволяют заключить, что данный многомодальный интерфейс обеспечивает приемлемую скорость и производительность работы пользователя с компьютером, не сильно отличающуюся от аналогичных показателей для стандартных контактных интерфейсов — устройств ввода, и может успешно применяться для бесконтактного управления как обычными операторами, так и потенциальными пользователями-инвалидами с грубыми моторными нарушениями в функционировании рук и даже вовсе без верхних конечностей.

Применение ассистивного пользовательского интерфейса позволит повысить социоэкономическую интеграцию инвалидов в информационном обществе и сделает их более независимыми от помощи со стороны других лиц. Предложенный бесконтактный интерфейс позволит пользователям самим выбирать доступные им средства взаимодействия с компьютером, компенсируя недоступные модальности альтернативными коммуникативными каналами.

Литература

1. *Карпов А. А.* ICanDo: Интеллектуальный помощник для пользователей с ограниченными физическими возможностями // Вестник компьютерных и информационных технологий, 2007. № 7. С. 32–41.
2. *Кричевец А.* Шлемомышь // Компьютерра, 2002. № 434. С. 48–51. www.computerra.ru/offline/2002/434/16588/.
3. *Bates R., Istance H. O.* Why are eye mice unpopular? A detailed comparison of head and eye controlled assistive technology pointing devices // 1st Cambridge Workshop on Universal Access and Assistive Technology Proceedings. — USA, 2002.
4. *Аграновский А. В., Евреинов Г. Е., Яшкин А. С.* Аппаратно-программные инструментальные средства проектирования виртуальных акустических объектов и сцен для слепых пользователей персональных компьютеров // Информационные технологии в образовании: Мат-лы IX Междунар. конф.-выставки. — М., 1999.
5. *Карпов А. А., Ронжин А. Л.* Многомодальные интерфейсы в автоматизированных системах управления // Известия высших учебных заведений. Приборостроение, 2005. Т. 48. № 7. С. 9–14.
6. *Ронжин А. Л., Карпов А. А.* Проектирование интерактивных приложений с многомодальным интерфейсом // Докл. Томского гос. ун-та систем управления и радиоэлектроники (ТУСУР), 2010. № 1. Ч. 1. С. 124–127.
7. *Ward D., Blackwell A., MacKay D.* Dasher: A data entry interface using continuous gestures and language models // ACM Symposium on User Interface Software and Technology UIST'2000 Proceedings. — New York: ACM Press, 2000. P. 129–137.
8. *Ronzhin A. L., Karpov A. A.* Russian voice interface // Pattern Recognition and Image Analysis (Advances in Mathematical Theory and Applications), 2007. Т. 17. № 2. С. 321–336.
9. *Карпов А. А.* Аудиовизуальный речевой интерфейс для систем управления и оповещения // Известия Южного федерального ун-та. Технические науки, 2010. № 3(104). С. 218–222.
10. *Lucas B. D., Kanade T.* An iterative image registration technique with an application to stereo vision // 7th Joint Conference (International) on Artificial Intelligence IJCAI Proceedings. — Vancouver, Canada, 1981. P. 674–679.
11. *Bouguet J.-Y.* Pyramidal implementation of the Lucas–Kanade feature tracker description of the algorithm // Intel Corporation Microprocessor Research Labs: Report. — New York, USA, 2000.

12. *Viola P., Jones M.* Rapid object detection using a boosted cascade of simple features // IEEE Conference (International) on Computer Vision and Pattern Recognition Conference (CVPR) Proceedings. — Kauai, HI, USA, 2001.
13. *Lienhart R., Maydt J.* An extended set of Haar-like features for rapid object detection // IEEE Conference (International) on Image Processing (ICIP'2002) Proceedings. — Rochester, New York, USA, 2002. P. 900–903.
14. *Gorodnichy D., Roth G.* Nouse 'Use your nose as a mouse' perceptual vision technology for hands-free games and interfaces // Image and Vision Computing, 2004. Vol. 22. No. 12. P. 931–942.
15. ISO 9241-9:2000(E) Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs). Part 9: Requirements for Non-Keyboard Input Devices. — International Standards Organization, 2000.
16. *Soukoreff R. W., MacKenzie I. S.* Towards a standard for pointing device evaluation, perspectives on 27 years of Fitts' law research in HCI // Intern. J. Human Computer Studies, 2004. Vol. 61. No. 6. P. 751–789.
17. *Zhang X., MacKenzie I. S.* Evaluating eye tracking with ISO 9241 Part 9 // Human–Computer Interaction Conference (International) (HCII 2007) Proceedings. — Beijing, China: Springer Verlag LNCS 4552, 2007. P. 779–788.
18. *Carbini S., Viallet J. E.* Evaluation of contactless multimodal pointing devices // 2nd IASTED Conference (International) on Human–Computer Interaction Proceedings. — Chamonix, France, 2006. P. 226–231.
19. *De Silva G. C., Lyons M. J., Kawato S., Tetsutani N.* Human factors evaluation of a vision-based facial gesture interface // Workshop on Computer Vision and Pattern Recognition for Computer Human Interaction Proceedings. — Madison, USA, 2003.
20. *Wilson A., Cutrell E.* FlowMouse: A computer vision-based pointing and gesture input device // Human–Computer Interaction INTERACT Conference Proceedings. — Rome, Italy, 2005. P. 565–578.
21. Видеодемонстрации с интернет-сайта лаборатории речевых и многомодальных интерфейсов СПИИРАН. www.spiiras.nw.ru/speech/demo/demo_new.avi, www.spiiras.nw.ru/speech/demo/ort.avi.

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2012. Т. 6. Вып. 2. С. 87–96

ЛОГИКА БИОГРАФИЧЕСКИХ ФАКТОВ

Н. А. Маркова[1]

Аннотация: Предложен метод формализации биографических фактов в виде логических формул, который позволяет интегрировать и анализировать данные, получаемые из разных источников, а также служит основой повышения эффективности справочно-информационного аппарата биографических ресурсов.

Ключевые слова: биографическое исследование; информационный поиск; формализация; биографический факт

1 Введение

Использование информационных технологий для проведения биографических исследований играет важную роль для удовлетворения объективно существующей общественной потребности в изучении и распространении сведений биографического характера. А потребность эта растет по мере роста общественного интереса к различным вариантам «частной» истории, в том числе истории науки, краеведения, истории семьи.

Колоссальный объем биографических сведений хранится в архивах или опубликован в труднодоступных изданиях. Массовая электронная публикация источников существенно повышает их доступность, однако потенциал информационных технологий по обеспечению эффективности биографического поиска далеко не исчерпан. Имеющиеся биографические ресурсы (БР) плохо систематизированы, их средства поиска неэффективны, данные противоречивы, а часто и недостоверны. Проблемы эти (подробно рассмотренные в работах [1, 2]) усугубляются, когда в процессе исследования объединяются данные, получаемые из разных источников. Для того чтобы процесс исследования был эффективен, требуется методическая и инструментальная поддержка как со стороны справочно-информационного аппарата БР, так и со стороны организации и систематизации работы исследователя.

В работе предложен способ представления формализуемых биографических данных в виде логических формул, который, с одной стороны — стороны исследователя, позволяет интегрировать данные, получаемые из разных источников, проверять непротиворечивость, интерполировать, корректно ставить новые исследовательские вопросы. С другой стороны — стороны БР, может служить концептуальной основой для «стандартов операционной совместимости, метаданных, средств упорядочения информационного содержания, интерфейсов доступа к массивам данных в цифровом формате, средств поиска и средств сохранения» — достижения целей, определяемых программой ЮНЕСКО «Информация для всех» [3].

В разд. 2 уточняются задачи биографического поиска, являющиеся важным звеном в исторических исследованиях самой разной направленности. В разд. 3 рассматриваются существующие модели биографических данных, предназначенные для систематизации и упорядочения сведений и обмена данными. В разд. 4 предлагаются основные концепции новой модели, конкретизируемые в разд. 5 в виде формул логики биографических фактов. В разд. 6 основные операции биографического поиска представлены в терминах формул логики биографических фактов.

2 Задачи биографического поиска

Задача биографа — собрать и обобщить биографические факты, под которыми понимаются высказывания, являющиеся «ответом на вопросы типа кто? что? когда?» [4, с. 53], упорядочить их определенным образом, связать между собой и с внешними объектами. Задача обобщения — интеграции данных — включает анализ их непротиворечивости и разрешение противоречий, интерполяцию, экстраполяцию, что может приводить к постановке новых исследовательских вопросов.

Объект исследования биографического поиска — это человек или группа людей, принадлежащих к определенному кругу, информация о которых сохранилась в источниках.

Предмет исследования биографического поиска — это конкретные биографические характеристики, отношения и события, связанные с изучаемыми людьми.

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

Биографический поиск — это задача, начинающаяся со слова «найти», которая может быть как независимой, так и включаться в виде определенного этапа в те или иные исторические исследования. В рамках предлагаемого рассмотрения в биографический поиск не включаются задачи причинно-следственного анализа исторических явлений, теоретических обобщений и художественных построений.

Эффективность биографических исследований во многом определяется доступностью информации, в том числе предоставляемыми БР возможностями поиска и навигации, а также наглядностью представления найденных данных. В рамках традиционных бумажных БР задача повышения доступности решается библиографами, архивистами и редакторами. Сведения об основных лицах, относящихся к документам, содержатся (во всяком случае, должны содержаться) в каталогах библиотек и описях архивов. Сведения о многочисленных лицах, упоминаемых в монографиях, — в соответствующих именных указателях. Библиографические справочники сопоставляют именам (сопровождаемым, возможно, краткими сведениями о лицах) указания на источники, в которых может быть найдена соответствующая информация.

Если направление деятельности библиографа — от источника к персонажам, то исследователь идет от персонажа к источникам и фактам и от объекта или явления — к персонажам. В качестве объекта или явления могут выступать как организации, общества, исторические события, так и лица, связи с которыми изучаются («корреспонденты Гоголя», «учителя Пушкина», «ученики Ключевского»). Аналогичные задачи стоят в рамках изучения истории любой сферы деятельности, отрасли, организации, края. В задачах генеалогии (или шире — истории семьи) изучаются как объекты «род», «семья», так и отдельные лица.

Отметим взаимосвязь описываемых сторон: для описания источников требуется выполнить задачу биографического поиска, а публикация результатов исследований приводит к созданию нового источника.

Биографический поиск — это многошаговая процедура. На основании исходных данных ставится некий исследовательский вопрос, ищутся источники, анализируются найденные документы, выявляются факты (или констатируется, что источник не содержит релевантных фактов), которые затем сопоставляются и интегрируются с ранее обнаруженными. На основе новой информации формулируются новые вопросы и т. д. Многие задачи и проблемы биографического поиска инвариантны по отношению к виду исследований.

3 Существующие биографические модели

Определить в общем случае, что должно входить в биографию, невозможно, однако в рамках проблемных областей, где возможна частичная формализация, такая задача имеет практические решения: от рекомендаций по подготовке повествовательных текстов до строгих стандартов представления данных. Рассмотрим основные виды биографических моделей, отмечая их достоинства и недостатки с точки зрения задач биографического поиска.

Упорядочение работы авторов биографического словаря определяется наличием рекомендаций по содержанию и форме представления данных. Например, грандиозная работа по созданию биографического словаря русских фольклористов методически обеспечена монументальным трудом [5], в котором наряду со специфическими рекомендациями есть и правила представления общих биографических и связанных с ними библиографических сведений. Такого рода БР рассчитаны на бумажную публикацию, в них нет учета возможностей информационных технологий.

Расширение горизонта рассмотрения от конкретной сферы деятельности до всех значимых в истории лиц вместе с освоением новых технологических возможностей (*wiki*), с одной стороны, и подключением широчайшего круга авторов, с другой, демонстрирует портал Персоналии в Википедии [6]. Методическим обеспечением для авторов-составителей статей служат шаблоны — своего рода биографические модели, каждая из которых ориентирована на некоторый круг лиц. Например, шаблон «Ученый» представляет собой список из двух десятков анкетных статей, каждая из которых раскрывается отдельным шаблоном-регламентом. Возможность ссылаться (гиперссылками) на другие статьи, относящиеся как к людям, так и к другим объектам, а также к доступным в сети источникам — колоссальное преимущество Википедии. Часть биографических данных можно почерпнуть в связанных статьях. *Wiki*-технология дает возможность совместно представлять формализованное и неформализованное знание. При этом формализация (в отличие от реляционных баз данных) может осуществляться по ходу накопления данных: подключением новых или изменением старых шаблонов. Несовершенства Википедии, возможно, частично будут сниматься по мере ее развития. Отметим некоторые из них.

Связь между статьями дается простой гиперссылкой, целесообразно расширить ее некоторым

семантическим содержанием (пометкой «друг», «отец»), что уже предполагают современные микроформаты [7]. Другое необходимое расширение, которое, к сожалению, пока даже не декларируют создатели семантического *Web*, — это определение динамики связей. Привязка событий и связей в жизни человека к временно́й оси — важнейшее условие успешности биографического поиска.

Важным шагом в сторону эффективного представления биографических данных является новая редакция Российского коммуникативного формата представления авторитетных данных в машиночитаемой форме (RUSMARC) [8].

Для лиц, причастных к созданию документов или упоминаемых в них, записи RUSMARC определяют имя и идентифицирующие признаки (даты жизни, специальность, область деятельности, титулы, звания, степени и т. п.). Аналогичные сведения полагаются для объектов библиографического описания «Род» (семья) и «Организация». Предполагается фиксация связей по родству, работе, культурной общности, местожительству и т. п. Как и Википедия, RUSMARC позволяет фиксировать связи между людьми и объектами другой природы. И эти связи также, к сожалению, не помечаются хронологическими рамками. Большое внимание в RUSMARC уделяется вариативности в именованиях лиц, организаций, документов — типичной причины проблем биографического поиска.

Основная сложность внедрения RUSMARC — отсутствие необходимого числа библиографов, которые могли бы заполнить соответствующие записи, в то время как успех Википедии во многом определяется подключением широчайшего круга лиц к созданию, проверке, редактированию статей.

Стандартом де-факто для представления биографической информации на протяжении долгих лет является давно устаревшая модель Genealogical Data Communications (GEDCOM), революционное, основанное на *xml* обновление которой GEDCOM 6.0 [9] было выпущено в 2002 г., но до сих пор фактически никем не используется. Конкурентом GEDCOM, тоже определяемым как спецификация формата обмена данными между генеалогическими программами, является стандарт GenXML [10]. Помимо более строгой структурной упорядоченности он несет в себе несколько принципиально новых положений, отражающих практику биографических исследований. В частности, в нем явным образом определяется процесс исследования: введены понятия «свидетельство» и «заключение». Но главное, GenXML открыт для добавления новых типов атрибутов и событий. К сожалению, ни в GEDCOM, ни в GenXML не отражены важнейшие свойства биографической информации: временна́я изменчивость и взаимная зависимость характеристик.

В рамках просопографических исследований [11] задача построения обобщенной модели и не ставится. Каждое исследование предполагает свою проблемно-ориентированную информационную модель.

Таким образом, в настоящее время концептуальных моделей, в полном объеме отражающих специфику биографических исследований, не существует.

4 Концептуальная модель биографических данных

Построим концептуальную модель, описывающую биографические данные, для которых возможно формализованное представление. Постараемся учесть все недочеты и достоинства существующих моделей. Дадим общее неформальное описание проблемной области биографических исследований.

Человек рождается, умирает, действует сам или подвергается воздействиям окружения в исторической реальности. Некоторая часть сведений о нем фиксируется документально и попадает в информационное пространство (ИП) — на бумажные и другие твердые носители, а в последние десятилетия и в электронные ресурсы. Информационное пространство является компонентом исторической реальности. Выделенные компоненты ИП — биографии конкретных лиц — существенно различаются по объему: от нескольких слов до нескольких томов. Кроме них биографические сведения рассыпаны по ИП — они содержатся в биографиях лиц из круга общения, в исторических описаниях событий и явлений, в документах учета, в библиографических списках и т. д. Задача биографического поиска — собрать рассыпанные в ИП данные, касающиеся конкретных лиц, конкретных событий, объектов, явлений.

Уточним вводимые понятия и термины.

4.1 Историческая реальность

С биографической точки зрения объектами исторической реальности (b-объектами) являются:

- люди (персонажи, лица);
- общественные образования (государства, учреждения, общества и др.);
- физические объекты (географические, технические, естественные);

– исторические события и процессы;

– отрасли деятельности.

Между b-объектами существует объективно или могут быть определены в рамках той или иной интерпретации различные отношения (b-отношения). Объективны биологические b-отношения, например отношение «ребенок–родители». Влияние на творчество писателя произведений его предшественника — пример субъективно интерпретируемого b-отношения.

Как выделение b-объектов, так и определение их свойств и b-отношений является результатом абстрагирования, необходимого для целей сбора и систематизации данных. Будем рассматривать только данные, для которых возможно формализованное представление. Интерпретация тонких вопросов, связанных с психологией, этикой, мировоззрением, творчеством, — задача соответствующих профессионалов.

Все b-объекты, а также большинство b-отношений существуют и изменяются во времени. Собственно «биографией» является информационный объект, в котором в динамике или интегрально представлены свойства и характеристики некоторого лица, а также его b-отношения с другими b-объектами и в какой-то мере их свойства и характеристики. Характеристики b-объекта или b-отношения будем называть b-характеристиками. Личными b-характеристиками являются составляющие генотипа и фенотипа человека, в частности состояние здоровья. Гражданское состояние, имущественное состояние, сословие, чин, звание, сан, титул характеризуют не только человека, но и соответствующую социально-правовую организацию общества, а точнее — b-отношение между ними.

Примером b-отношения, на первый взгляд не зависящего от исторического контекста, является отношение местопребывания, связывающее человека и объект географического пространства. Формально в конкретный момент времени местопребывание может быть охарактеризовано координатами. Однако на практике оно определяется в терминах названий населенных пунктов, сопоставление которых с координатами — не всегда тривиальная задача исторической географии.

Важнейшая b-характеристика — официальное именование — определяется на отношении лица и b-объекта-государства. Другие b-объекты могут использовать другие имена данного лица, в частности в домашнем обращении или в отношении авторства (псевдонимы).

Между b-характеристиками существуют зависимости, регламентируемые законами природы или нормативными законами: правилами, юридическими актами, традициями. Примерами регламентов являются законодательные документы, уставы обществ, штатное расписание учреждения и т. п. Существуют закономерности, определяющие допустимые последовательности событий — смены значений b-характеристик, задающие некий шаблон, сценарий или набор ограничений. Перечислим несколько очевидных: смерть следует за рождением; имеются допустимые пределы разницы между рождением человека и границами жизни его родителей; в каждый конкретный момент времени человек может находиться только в одной точке географического пространства. Для определения большинства нормативных законов требуется конкретно-историческое знание.

4.2 Информационное пространство

Информационное пространство без ограничения общности можно представить как совокупность хранилищ документов. Для пользователей электронных хранилищ, реализованных, возможно, в виде баз данных, их содержание представляют виртуальные документы, визуализируемые на экране. Документы, в свою очередь, делятся на части/фрагменты: разделы, страницы, абзацы и т. п. Фрагмент документа редко независим, для его корректной интерпретации требуется контекст. Элементы ИП также являются объектами исторической реальности: у них есть время жизни, они связаны с людьми отношениями «автор», «адресат», «упоминаемое лицо».

Документы, их фрагменты или их совокупности — хранилища, и их разделы идентифицируются адресами. Для электронных хранилищ адрес — это URL/URI (Uniform Resource Locator/Identifier), ключи базы данных, имена закладок и т. п. Для архивных хранилищ — номера фонда, описи, дела, листа. Сложнее дело обстоит с печатными изданиями. Библиографическая ссылка, вообще говоря, не является адресом — книгу еще требуется найти в хранилище-библиотеке, где адресом ее будут служить соответствующие шифры хранения. Возможно, однако, что книга уже оцифрована, тогда адрес ее — тот же URL.

Идеальное решение задач биографического поиска состояло бы во всеобщем справочно-информационном аппарате, в котором b-объекты были бы соотнесены с элементами ИП. На базе такого аппарата, используя возможности современных информационных технологий, удалось бы добиться качественно нового уровня эффективности биографических исследований. На пути к этому идеалу стоят сложнейшие задачи.

Прежде всего, задача описания существующих источников (как бумажных, так и электронных) — идентификации взаимоотношений между элементами ИП и b-объектами, их свойствами, хронологией — далека от реализации. «Концепция информатизации архивного дела России» [12], декларируя необходимость обеспечения прав граждан на информацию, на самом деле только намечает подходы к решению задач научного описания архивных материалов. В практике отечественных архивов электронные описания в основном присутствуют разве что на уровне их крупных единиц — фондов.

Существуют два аспекта задачи описания источников: систематизация и наполнение. В части наполнения многое могли бы сделать пользователи, читатели, исследователи. Примером деятельности по обмену биографическими ссылками (и фактографическими данными) может служить сайт ВГД [13]. Пользователь, обладающий доступом к труднодоступному источнику, выкладывает его описание на общедоступный ресурс. Но эта коммуникация ведется бессистемно: в виде обмена текстовыми сообщениями на форуме сайта.

Необходимым условием создания справочно-информационного аппарата, соотносящего элементы ИП с b-объектами, является систематизация их описаний, сведение формализуемой части касающихся их сведений в единый формат метаданных.

5 Биографические характеристики и логика фактов

Для того чтобы иметь возможность сопоставить факты, содержащиеся в источниках, оценить их непротиворечивость, сделать выводы, необходимо привести их к некоторому общему, нормализованному представлению. Представим биографические сведения в виде совокупности взаимосвязанных значений b-характеристик на общей временно́й оси.

5.1 Биографические характеристики

Существенная часть b-характеристик формализуема, их можно измерить, например, в терминах социологической [14] или психологической [15] стратификации. Значения b-характеристик — статусы (атрибуты), как правило, изменяются во времени. Например, отношение сотрудника и учреждения характеризует должность, которая изменяется по мере карьерного роста. В некоторых случаях новое значение не заменяет предыдущее, а присоединяется к списку ранее имевшихся компонентов значения (пример — награды). Ряд b-характеристик имеет простые количественные значения, например рост и вес. Артериальное давление измеряется парой чисел.

Значения b-характеристик в именных шкалах, где присутствуют синонимы, многозначны. Существует вариативность именования лиц и организаций, даже если речь идет о конкретном моменте времени. Например, «собор Покрова Пресвятой Богородицы, что на Рву» эквивалентен «собору Василия Блаженного». Причиной многозначности могут быть также те или иные варианты искажений, а также в целом субъективный характер оценок (про рост профессора Ловецкого Герцен утверждал: «был высокий. . . мужчина», а Пирогов — «небольшого роста»).

B-характеристика — это отображение (в общем случае многозначная функция), динамически сопоставляющее b-объекту или паре b-объектов некоторое значение. Область определения b-характеристики соответствует некоторому b-отношению.

Одно и то же b-отношение может быть оценено разными, но взаимозависимыми b-характеристиками в разных шкалах, например рост измеряется в саженях, футах или сантиметрах (или же ему дается неформальная словесная оценка). Кроме того, между значениями различных b-характеристик связанных между собой b-объектов существует взаимосвязь, в частности, являясь сотрудником подразделения, человек автоматически является и сотрудником учреждения в целом. С другой стороны, являясь сотрудником учреждения, человек работает в некотором подразделении, о котором, если оно неизвестно, может быть поставлен исследовательский вопрос.

Как области определения, так и области значений для подавляющего большинства b-характеристик меняются во времени, само наличие b-характеристики ограничено определенными временны́ми рамками, для конкретной исторической ситуации их определяет научное знание соответствующей специальной исторической дисциплины, а также социологии, антропологии, психологии, биологии. Они же определяют зависимости между b-характеристиками, варианты возможной синонимии значений и другие общие для рассматриваемых классов b-объектов закономерности. Формулировки соответствующих законов природы, нормативных актов, традиций в виде зависимостей между значениями b-характеристик будем называть b-нормалями. Лишь в редких случаях b-нормаль может быть сформулирована формально, большинство из них задается неформальными текстами, проверка соблюдения их правил — «ручная» процедура.

5.2 Нормализованный факт

Предложим формализацию понятия b-характеристики. Высказывание, фиксирующее, что в данный момент данная b-характеристика для данного b-объекта (объектов) имела данное значение, назовем формализованным фактом. Такое выражение может иметь логическое значение — ИСТИНА или ЛОЖЬ, быть неизвестным, а может быть оценено неким промежуточным образом: «Скорее, ИСТИНА, чем ЛОЖЬ». Если информация об интересующем b-объекте или группе объектов будет представлена в виде набора формализованных фактов, тогда, применяя соответствующие b-нормали, можно их сопоставить, выявить и разрешить противоречия, сформировать новые факты-следствия, интегрировать данные в общую картину.

Большинство b-характеристик сохраняет свои значения на протяжении некоторого промежутка времени. Чтобы отразить это фундаментальное свойство исторической реальности, введем понятие нормализованный факт (НФ). Назовем нормализованным фактом следующую логическую формулу:

$$(\forall t \in \Delta t)\, \beta(p, q, t) = a\,. \qquad (1)$$

Здесь β — b-характеристика; p и q — b-объекты; $a \in \in \mathrm{Im}\,(\beta)$ — конкретное значение b-характеристики из области ее значений; t — время; Δt — период времени, когда b-характеристика неизменна.

Компоненты НФ полагаем некоторым информационным представлением соответствующих сущностей, например в виде идентификаторов, текстов, чисел.

В формуле (1) представлена характеристика двуместного b-отношения. Для одноместных b-отношений будем использовать нотацию $\beta(p, t)$. Трехместные отношения, а также отношения большей местности с помощью логических формул могут быть сведены к двуместным.

Если b-характеристика принимает логическое значение, будем опускать его значение в записи, полагая $\beta(p, q, t)$ эквивалентным

$$\beta(p, q, t) = \text{ИСТИНА}\,.$$

Наконец, для краткости в формуле (1) будем опускать время или использовать нотацию

$$\beta(p, q, \Delta t) = a\,.$$

5.3 Формулы логики фактов

Расширим понятие НФ для различных предикатов. Помимо «=» в формуле (1) будем использовать «$\neq$»; «$<$» и «$>$» — для упорядоченных значений b-характеристик; а также «$\in$» и «$\notin$» — если в правой части не единичное значение, а множество. Кроме того, правая часть может представлять значение b-характеристики для другого b-объекта, что соответствует текстам: «был одноклассником», «служил в той же должности», «был старше чином».

Нормализованные факты связаны друг с другом. Формально эти связи представимы в виде логических формул. Назовем их формулами логики фактов — FF (*Fact Formula*). Для сигнатуры формул логики фактов применимы как выражения обычной логики предикатов, так и специальные темпоральные аппараты. В любом случае FF включает пропозициональные связки ($\neg$, $\neq$, $\vee$, $\rightarrow$) и кванторы ($\forall$, $\exists$). B-характеристики выступают в роли функций, а в качестве предикатов используются b-характеристики с логическим значением или оценки значений b-характеристик ($=$, $\neq$, $\in$, $\notin$, $<$, $>$, . . .).

Дадим индуктивное определение формулы логики фактов — FF в терминах логики предикатов.

Переменными и константами являются b-объекты (p, q), элементы и подмножества из множеств значений b-характеристик ($a \in \mathrm{Im}\,(\beta)$, $A \subset \mathrm{Im}\,(\beta)$) и время (t), а также различные варианты временны́х периодов (отрезок, интервал, полуинтервал):

$$\begin{aligned}
&\mathrm{Term} ::= \beta(p, q, t) | a | A | t | \\
&\qquad |[t_1 - t_2] | (t_1 - t_2) | [t_1 - t_2) | (t_1 - t_2]\,; \\
&\mathrm{Atom} ::= \mathrm{Term}\ \rho\ \mathrm{Term}\ \ (\rho \in \{=, \neq, \in, \notin, <, >, \ldots\}) \\
&\quad \mathrm{FF} ::= \mathrm{Atom} | \neg \mathrm{FF} | \\
&\qquad |\mathrm{FF}_1 \wedge \mathrm{FF}_2 | \mathrm{FF}_1 \vee \mathrm{FF}_2 | \forall x\ \mathrm{FF} | \exists x\ \mathrm{FF}\,.
\end{aligned}$$

Приведем формулировки некоторых b-нормалей в терминах логики фактов.

Симметрия. Для большинства двуместных b-отношений значению $\beta(p, q, t)$ однозначно соответствует $\beta'(q, p, t)$:

МестоРаботы(«Иванов», «Контора») $\leftrightarrow$

$\leftrightarrow$ *Сотрудник*(«Контора», «Иванов»)\,.

Транзитивность. Факты, основанные на таких b-характеристиках, как иерархия и местоположение, обладают свойством транзитивности. Например:

МестоРаботы(«Иванов», «Контора») $\wedge$

$\wedge$ *Местопребывание*(«Контора», «Москва») $\rightarrow$

$\rightarrow$ *Местопребывание*(«Иванов», «Москва»).

Вариативность. Для вариативных b-характеристик, например именования, значения разбиваются на классы эквивалентности, определяемые b-нормалями, а формула принимает вид принадлежности данному классу. Приведем b-нормаль для имено-

вания в современной отечественной практике (без учета знаков препинания и грамматических форм, в нестрогой форме):

Имя(x) = «Имя» $\wedge$
$\wedge$ Фамилия(x) = «Фамилия» $\wedge$
$\wedge$ Отчество(x) = «Отчество» $\rightarrow$
$\rightarrow$ Именование(x) = «Фамилия Имя Отчество» $\vee$
$\vee$ «Имя Отчество Фамилия» $\vee$
$\vee$ «Фамилия Имя» $\vee$
$\vee$ «Имя Фамилия» $\vee$
$\vee$ «Имя Отчество» $\vee$
$\vee$ «И О Фамилия» $\vee$
$\vee$ «Фамилия И О» $\vee$
$\vee$ «И Фамилия» $\vee$
$\vee$ «Фамилия И» .

5.4 Виды нормализованных фактов

Рассмотрим два направления классификации НФ: по динамике и по определенности.

1. Нормализованные факты представляют различные варианты динамики значений b-характеристик:
   - НФ-событие: период соответствует «мгновению»: $\Delta t = \{t'\}$;
   - НФ-состояние: период протяжен — $\Delta t = [t_s, t_b)$, он включает начало и не включает конец временно́го промежутка;
   - цепочка НФ — процесс, дизъюнкция формул отдельных состояний: $\vee_{i\in[0,\,n-1]}\beta(p, q, [t_i,\, t_{i+1}))$ — в моменты t_i b-характеристика принимает новое значение.

   Факт-состояние — это элементарный процесс, включающий событие — переход в данное состояние и ограниченный событием — выходом из данного состояния. С другой стороны, интегральное состояние может уточняться детальным процессом.
2. Нормализованные факты представляют различные варианты определенности значений b-характеристик:
   - рамочный НФ — факт, часть компонентов которого не определена;
   - строгий НФ — факт, для которого все компоненты определены.

   Под компонентами факта понимаются время, характеризуемые объекты и значения b-характеристики.

   Заметим, что строгий факт совсем не обязательно корректен. Произвольная формула логики фактов является строгой, если все ее компоненты — строгие факты, и рамочной во всех других случаях.

Неполные и неточные данные, оформленные в виде рамочных НФ и основанных на них рамочных формул, пусть с пропусками, размытыми значениями, неформальными комментариями при накоплении, систематизации, анализе способны стать основой для реконструкции биографий.

6 Биографический поиск в терминах логики фактов

Рамочные НФ представляют удобный механизм для формулировки исследовательских вопросов, ответы на которые, возможно, хранятся в документах-источниках. Как интерпретация источников, так и выводы из имеющихся фактов, как правило, являются неформальными процедурами. Для их выполнения требуется экспертное знание. Тем не менее сам факт выполнения операции, ее вход и выход удобно представить как специальные формулы логики фактов. Благодаря формализованному представлению ручных операций сведения о целях исследования, его текущем состоянии и дальнейших шагах складываются в единую картину.

6.1 Формулировка исследовательских вопросов

Рамочный НФ связан с исследовательским вопросом, касающимся уточнения значений его компонентов. Рассмотрим основные варианты таких вопросов.

Хронологические рамки. Важнейший вопрос биографического поиска — «когда?». Например, Жизнь(«Иванов», $[t_1 - t_2)$), где t_1 и t_2 неизвестны. При его постановке, как правило, существуют ориентиры, оценки: начало не ранее, конец не позднее — или известен некоторый промежуток, входящий в искомый.

Рамки значений. Исследовательский вопрос состоит в выяснении, какое значение принимала данная b-характеристика для данного лица в данное время. Пример:

$$\textit{СотрудникДолжность}(\text{«Контора»},\ \text{«Иванов»}, 1890) = x.$$

Область значений:

$$x \in \textit{СписокДолжностей}(\text{«Контора»}, 1890)\,.$$

Уточнение рамок возможно, если известны значения b-характеристики в некоторые моменты до и после данного времени и, кроме того, ее значения упорядочены или известно значение другой b-характеристики, а между b-характеристиками существует зависимость, определяемая b-нормалью.

Объектные рамки. Неопределенность этого рода — ответ на вопросы «кто?» — ***Муж***$(x,$ «Петров», 1909) или «что?» — ***МестоРаботы***(«Иванов», y, 1890). Запись в анкете «женат», предполагает наличие b-объекта x — жены, про которую известно лишь то, что в указанное время она была замужем за заполнителем анкеты. Собственно поиск объекта сводится к поиску его b-характеристик, по крайней мере, идентификационных (именования, времени жизни). Какие-то ограничения на именование и время жизни содержат уже имеющиеся данные о связываемом b-объекте.

6.2 Интерпретация текстов источника

Лишь незначительная часть ИП представляет формализованные факты в явном виде. Это в основном фактографические (генеалогические или просопографические) базы данных, в которых b-характеристика представляется доменом (колонкой в таблице), а шкала ее значений, возможно, выделена в отдельную вспомогательную таблицу-словарь.

Там, где документ-источник осмысленно размечен, например в анкетных группах — шаблонах Википедии, формализованные факты представлены метаданными.

Кроме того, несложно выявить формализованные факты, исходя из:

- структурной организацией документа-источника, отвечающей структуре b-объектов;
- в той или иной степени структурированных текстовых определений, сопоставляющих именам b-объектов имена связанных с ними b-объектов или b-характеристики;
- идеографических схем (пример — генеалогическое древо), определяющих связи между b-объектами.

Структура справочного издания «Памятная книга губернии» [16] отражает деление губернии на уезды и населенные пункты, принадлежность учреждений как административным единицам, так и ведомствам, структуру учреждений и должности, звания, чины служащих в них лиц.

Однако большинство фактов извлечь из текстов может только исследователь. Определим операцию интерпретации ($\Rightarrow$), недоступную или неэффективную для автоматического логического анализа и выполняемую исследователем, которая выводит формулы логики фактов из содержания компонента ИП. Например, то, что в тексте речь идет об указанном промежутке времени, представимо следующей формулой: $(text) \Rightarrow \exists t \in [t_{\min}, t_{\max}](text)$.

Другой пример — текст содержит сведения о b-характеристике β и некотором подмножестве ее значений A: $(text) \Rightarrow \exists x(\beta(x) \in A)(text)$. В частности, если текст относится к b-объекту, именуемому «NN»: $(text) \Rightarrow \exists x($***Именование***$(x) = $ «NN»$)(text)$.

Максимально точная оценка биографических границ для текста, реализуемая как внутренняя разметка или как внешнее индексирование, должна способствовать эффективности биографического поиска.

Интерпретация источника может выявить новые неизвестные b-отношения, которые на новом шаге позволят найти дополнительные сведения об искомом b-объекте. Например, в воспоминаниях одноклассника искомого лица мы находим сведения об их учителе (неизвестное ранее b-отношение), а уже в переписке учителя — сведения об искомом лице.

Немаловажный факт, извлекаемый из текста, состоит в утверждении, что в нем нет сведений о данном b-объекте или о данной b-характеристике.

6.3 Выборка биографических сведений из источника

Выборку текста так же, как его интерпретацию, представим в терминах логики фактов. Введем следующее обозначение: $text =$ ***Контент***(l) — интерпретируемый текст — это содержание документа (фрагмента документа) по адресу l. Расширим множество констант и переменных логики фактов произвольными текстами и адресами в ИП.

Важнейшим фактом, извлекаемым из текста, является отсылка к другим компонентам ИП. Пусть из интерпретации текста по адресу l_0 следует, что сведения относительно b-характеристики β b-объекта x можно найти в документе по адресу l_1, тогда: ***Контент***$(l_0) \Rightarrow$ ***Контент***$(l_1) \wedge \beta(x)$.

Доступ к архивным материалам или редким изданиям затруднен. В то же время необходимый для анализа в конкретном исследовании объем данных, как правило, бывает много меньше документа-источника в целом. Поэтому как в случае труднодоступных источников, так и для анализа легкодоступных, но объемных материалов, применяется практика создания вторичных документов: копии фрагмента, выписки, реферата, аннотации. В терминах логики фактов то, что по отношению к совокупности исследовательских вопросов (ff)

при *Редукции* — создании вторичного документа — не выпущено ничего существенного, фиксируется так:

$$(\textit{Контент}(l) \wedge \text{ff}) = (\textit{Редукция}(\textit{Контент}(l)) \wedge \text{ff})\,.$$

6.4 Вывод формул логики фактов

Появление новой формулы логики фактов в багаже исследователя возможно и как следствие интерпретации некоторого текста, и как логический вывод из имеющихся формул. И в том, и в другом случае возможны неточности и огрехи, вследствие как ошибок исследователя, так и некорректности или неполноты исходных данных. В любом случае полезно делать предположения, строить рабочие гипотезы. Для того чтобы отличить гипотезу от установленного факта, имеет смысл воспользоваться категориями нечеткой логики, т. е. вместо значений ИСТИНА и ЛОЖЬ, употреблять оценку правдоподобия в виде числа в диапазоне от 0 (ЛОЖЬ) до 1 (ИСТИНА). В виде формул нечеткой логики могут быть сформулированы опирающиеся на статистику экспертные оценки. Пример:

$$\begin{aligned}
&\textit{Оценка}(\textit{Служба}(\text{«Леонтий Кириллович»},\\
&\text{«церковь села Ловцы»}, 1780{-}1790) \wedge\\
&\wedge\, \textit{Служба}(\text{«Кирилл Васильевич»},\\
&\text{«церковь села Ловцы»}, 1755{-}1760) \wedge\\
&\wedge\, \textit{ЭкспертнаяОценка}(\textit{Имя}(\text{духовное лицо}) =\\
&=\ \text{«Кирилл»}) = 0{,}003 \wedge\\
&\wedge\, \textit{ЭкспертнаяОценка}((\textit{Служба}(x, c) \wedge\\
&\wedge\, \textit{Родня}(x, y)) \to \textit{Служба}(y, c)) = 0{,}75 \to\\
&\to \textit{Отец}(\text{«Леонтий Кириллович»},\\
&\text{«Кирилл Васильевич»})) = 0{,}9\,.
\end{aligned}$$

Нечеткую логику рационально использовать и для оценки противоречивых фактов. Нормализованный факт с оценкой, отличной от 0 и 1, представляет собой вариант исследовательского вопроса. Важно сохранять все, даже отвергнутые (с оценкой 0), факты. На новом этапе исследователь, может быть, вернется к ним, в том числе для анализа намеренных искажений, из которых также могут быть выявлены некоторые факты или дана предопределенная оценка извлекаемых из некорректного источника новых фактов.

7 Заключение

Подытожим статью, перечислив основные черты предложенной модели представления биографической информации:

- в рассмотрение включаются не только лица, но и объекты иной природы (организации, населенные пункты, исторические события) вместе с их структурой и историей;
- помимо характеристик отдельного объекта рассматриваются характеристики отношений между объектами;
- характеристики рассматриваются в динамике изменения их значений;
- наличие характеристик, их возможные значения и взаимозависимости определяются конкретно историческими знаниями, большая часть которых также изменяется во времени;
- для представления совокупности установленных фактов, намеченных к рассмотрению вопросов, а также шагов процесса исследования предложена единая форма — формулы логики биографических фактов.

Предложенная модель не противоречит существующим биографическим моделям, а, скорее, дополняет и уточняет их, расставляя несколько другие акценты. В частности, приоритет динамических характеристик отношений между объектами позволяет создать целостную непротиворечивую картину биографии отдельного лица и в то же время обеспечивает повторное использование фактов при освещении биографий связанных с ним лиц или истории объектов другой природы.

То, что номенклатура рассматриваемых характеристик, их возможные значения и допустимые отношения между ними не закреплены жестко, а определяются динамически изменяемыми нормалями, обеспечивает открытость модели, возможность применения ее для решения задач из самых разных проблемных областей.

На основе предложенной модели возможна организация эффективного доступа к БР. Для этого определение метаданных, с одной стороны, и интерфейса поисковых запросов — с другой, следует осуществлять в виде формул логики фактов. Такой подход применим не только при создании новых БР, но и для модификации существующих открытых БР, а также для разработки специализированных оболочек закрытых БР.

Важнейшим приложением модели должно стать создание рабочего поля биографического исследования — своеобразного «нового» БР, в котором востребованы возможности работы с неточными, противоречивыми, непроверенными данными. Их постепенное накопление, систематизация, выдвижение и опровержение гипотез, сопоставление, обоснование, постановка новых вопросов — все то, что необходимо в процессе исследования, удобно представимо с помощью аппарата логики фактов.

Литература

1. *Маркова Н. А., Адамович И. М.* Электронные биографические ресурсы // Электронные библиотеки: перспективные методы и технологии, электронные коллекции (RCDL'2010): Труды XII Всеросс. научн. конф. — Казань: КГУ, 2010. С. 168–180.
2. *Маркова Н. А., Адамович И. М.* Коллекции персоналий // Системы и средства информатики. Вып. 20. № 2. Методы и технологии, применяемые в научных исследованиях информатики. — М.: ИПИ РАН, 2010. С. 178–198.
3. Стратегический план Программы ЮНЕСКО «Информация для всех» (2008–2013 гг.). — М.: Межрегиональный центр библиотечного сотрудничества, 2009. 48 с.
4. *Валевский В. Л.* Биографика как дисциплина гуманитарного цикла // Лица: биографический альманах. — СПб.: Феникс, 1995. Вып. 6. С. 33–68.
5. Русские фольклористы: Биобиблиографический словарь. Пробный выпуск / Отв. ред. Т. Г. Иванова и А. Л. Топорков. — М.: ПРОБЕЛ-2000, 2010. 240 с.
6. Портал Персоналии. Материал из Википедии — свободной энциклопедии. http://ru.wikipedia.org/wiki/Портал:Персоналии (дата обращения: 29.01.2011).
7. Микроформаты. http://microformats.org/ (дата обращения: 29.01.2011).
8. Российский коммуникативный формат. — Министерство культуры Российской Федерации, Российская библиотечная ассоциация, Национальная служба развития системы форматов RUSMARC. http://www.rba.ru/rusmarc/.
9. GEDCOM XML Specification, Release 6.0. http://xml.coverpages.org/Gedcom-XMLv60.pdf (дата обращения: 29.01.2010).
10. GenXML 3.0 16.06.2010. http://www.cosoft.org/genxml/GenXML30.pdf (дата обращения: 29.01.2011).
11. *Юмашева Ю. Ю.* Историография просопографии // Известия Уральского государственного университета. — Екатеринбург: УрГУ, 2005. № 39. Гуманитарные науки. Вып. 10. С. 95–127.
12. Концепция информатизации архивного дела России. Утверждена Росархивом в 1995 г. http://www.rusarchives.ru/informatization/conseption.shtml (дата обращения: 29.01.2011).
13. Всероссийское генеалогическое древо (ВГД). http://baza.vgd.ru/ (дата обращения: 29.01.2011).
14. *Кравченко А. И.* Социология. Общий курс: Учебное пособие для вузов. — М.: ПЕРСЭ; Логос, 2002. 640 с.
15. *Ганзен В. А.* Системные описания в психологии. — Л.: ЛГУ, 1984. 175 с.
16. Памятные книжки губерний и областей Российской империи. http://www.nlr.ru/pro/inv/mem_buks.htm (дата обращения: 29.01.2011).

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2012. Т. 6. Вып. 2. С. 97–100

РАСЧЕТ И ОПТИМИЗАЦИЯ НЕКОТОРЫХ ХАРАКТЕРИСТИК ДЛЯ МОДЕЛИ ВЫЧИСЛИТЕЛЬНОГО КОМПЛЕКСА

И. В. Павлов[1]

Аннотация: Рассматривается проблема выбора оптимального размера пакетов при обработке информационных задач большого объема для модели вычислительного комплекса с учетом возможных отказов или сбоев элементов в процессе решения задачи. Получено приближенное асимптотическое решение данной проблемы для случая высоконадежных элементов и малого времени пересылки (загрузки) пакетов.

Ключевые слова: оптимальный размер пакета; надежность; интенсивность отказов; время пересылки пакетов

1 Введение

Пусть имеется система, включающая в себя l основных вычислительных элементов. В систему поступают «задания», каждое из которых состоит из некоторого (вообще говоря, случайного) числа «элементарных задач», каждая из которых может выполняться (обрабатываться) независимо от остальных на любом из этих элементов. Для выполнения очередного задания, поступившего в систему, необходимо выполнить все составляющие его элементарные задачи. При этом в процессе выполнения задание разбивается на некоторое количество n блоков («пакетов») элементарных задач равного объема $v = L/n, n \in N$, где N — множество допустимых значений (например, N — некоторое подмножество целочисленных значений, кратных 2 и т. п.). Время h выполнения одной элементарной задачи на любом из элементов далее будем считать равным единице: $h = 1$. Соответственно, время выполнения одного пакета объемом v на любом из элементов будет численно совпадать с величиной v.

Выполнение задания происходит путем пересылки пакетов на рабочие элементы и дальнейшей их обработки на этих элементах. Время пересылки (загрузки) пакета на элемент равно величине $\tau > 0$, не зависит от размера пакета v и от состояния других элементов. Обработка пакета после его загрузки на данном элементе занимает время τ и происходит независимо от состояния других элементов. После завершения обработки очередного пакета на том или ином элементе снова происходит его загрузка в течение времени τ следующим пакетом (из общей очереди всех пакетов данного задания) независимо от состояния (работы или загрузки) остальных элементов и т. д. Задание считается выполненным после выполнения (обработки) всех составляющих его пакетов. Близкие по смыслу модели и процессы рассматривались ранее в [1–5].

В процессе работы любой из элементов может отказывать с постоянной (не зависящей от времени) функцией интенсивности отказов $\lambda(t) \equiv \lambda$ [6, 7]. Заметим, что более близким к реальности было бы предположение о монотонном возрастании (неубывании) $\lambda(t)$ по времени. Поэтому фактически здесь предполагается, что, по крайней мере в течение времени выполнения одного задания, функция интенсивности отказов $\lambda(t)$ меняется незначительно и может считаться приближенно постоянной. Такое допущение является естественным, по крайней мере в случае высокой надежности элементов, когда вероятность отказа элемента за время выполнения в системе одного задания достаточно мала. В указанных допущениях время безотказной работы элемента имеет экспоненциальное распределение с функцией надежности $P(t) = e^{-\lambda t}$, а вероятность отказа элемента за время h выполнения одной элементарной задачи равна величине $\lambda h + o(\lambda h)$.

Одной из существенных проблем, возникающих в данной ситуации, является выбор оптимального размера пакета v с учетом возможности отказов (сбоев) элементов при выполнении задания.

2 Модель со сбоями элементов

Рассмотрим случай, когда возможные отказы элементов в системе имеют характер «сбоев». Другими словами, в результате отказа (сбоя) элемент сам по себе не выходит из строя и продолжает работать, но находящийся на нем в момент сбоя пакет считается невыполненным и после завершения его

[1] Московский государственный технический университет им. Н. Э. Баумана, ipavlov@bmstu.ru

обработки снова ставится в очередь необработанных пакетов и должен быть полностью обработан заново на этом же или любом другом элементе.

Рассмотрим сначала более простой частный случай, когда число элементов $l = 1$. Обозначим через $p = \exp(-\lambda\upsilon)$ вероятность обработки пакета объемом υ без сбоев и $q = 1 - p$. Время η выполнения всего задания объемом L имеет вид:

$$\eta = (\upsilon + \tau)\upsilon\,, \tag{1}$$

где υ — момент (номер шага) первого достижения n «успехов» в классической схеме независимых испытаний Бернулли при вероятности «успеха» (на одном шаге) $p = \exp(-\lambda\upsilon)$. Задача выбора оптимального размера пакета υ далее сводится к минимизации математического ожидания $\mathrm{E}\eta$ по параметру υ, или, учитывая равенство $\upsilon = L/n$, к минимизации $\mathrm{E}\eta$ по переменной $n \in N$, где n — число пакетов, на которое разбивается задание. Случайная величина υ имеет распределение Паскаля

$$P(\upsilon = m) = C_{m-1}^{n-1} p^n q^{m-n}\,, \quad m = n, n+1, \ldots, \tag{2}$$

с математическим ожиданием $\mathrm{E}\upsilon = n/p$, откуда с учетом (1) следует, что выбор оптимального размера пакета сводится к задаче: найти

$$\min\,(L + n\tau)\exp\left(\frac{\lambda L}{n}\right) \tag{3}$$

по $n \in N$. Далее оптимальный размер пакета $\tilde{\upsilon}$ находится по формуле $\tilde{\upsilon} = L/\tilde{n}$, где $\tilde{n} \in N$ — решение задачи (3).

Оптимизационная задача (3) является целочисленной, поскольку множество N допустимых значений n содержит только целочисленные точки. Введем также дополнительную «непрерывную» задачу: найти

$$\min\,(L + n\tau)\exp\left(\frac{\lambda L}{n}\right) \tag{4}$$

по всем (не только целочисленным) значениям $n \geq 1$. Далее оптимальный размер пакета υ^* (без ограничения целочисленности $n \in N$) находится как $\upsilon^* = L/n^*$, где n^* — решение задачи (4).

Теорема 1 дает точное решение оптимизационных задач (3) и (4). Теорема 2 дает асимптотическое выражение для оптимального размера пакета υ^*.

Теорема 1. *Пусть $\lambda > 0$, $\tau > 0$ и выполняется неравенство*

$$\tau \leq \lambda L^2\,. \tag{5}$$

Тогда минимум (4) *достигается в единственной точке*

$$n^* = L\sqrt{\frac{\lambda}{\tau}}\left[\sqrt{1 + \frac{\lambda\tau}{4}} + \frac{\sqrt{\lambda\tau}}{2}\right]\,.$$

Минимум (3) *достигается в одной из двух ближайших (слева или справа) к точке n^* целочисленных точек $n \in N$.*

Доказательство. Введем функцию

$$f(n) = (L + n\tau)\exp\left(\frac{\lambda L}{n}\right) \tag{6}$$

от непрерывного аргумента $n \geq 1$. Нетрудно показать, что знак производной этой функции совпадает со знаком многочлена $Q(n) = n^2 - \lambda L n - \lambda L^2/\tau$, который имеет при $n \geq 1$ единственный корень в точке $n = n^*$ и для которого справедливы неравенства:

$$Q(n) < 0 \ \text{ при } \ 1 \leq n \leq n^*\,;$$
$$Q(n) > 0 \ \text{ при } \ n > n^*\,,$$

если выполняется условие (5), откуда далее и следует теорема 1. Теорема доказана.

Теорема 2. *Пусть $\lambda > 0$, $\tau > 0$, $\tau \leq \lambda L^2$ и $\lambda\tau \to 0$. Тогда*

$$\upsilon^* = \sqrt{\frac{\tau}{\lambda}}\,[1 + o(1)]\,. \tag{7}$$

Доказательство следует из теоремы 1 и равенства $\upsilon^* = L/n^*$.

Из (7) далее следует приближенная формула для оптимального размера пакета при $\lambda\tau \ll 1$:

$$\upsilon^* \cong \sqrt{\tau\theta}\,,$$

где $\theta = 1/\lambda$ — математическое ожидание времени безотказной работы (средний ресурс) элемента. Другими словами, оптимальный размер пакета υ^* приближенно равен среднему геометрическому между временем пересылки (загрузки) τ и средним ресурсом элемента θ (при условии $\lambda\tau \ll 1$). Существенно, что оптимальное значение υ^* не зависит от размера всего задания L, который, вообще говоря, может быть неизвестным и случайным.

Рассмотрим далее общий случай $l \geq 1$ элементов. Для рассматриваемой модели время выполнения задания

$$\eta = (\upsilon + \tau)\left(\frac{\upsilon}{l}\right)^+\,, \tag{8}$$

где z^+ — величина z, округленная вверх до ближайшего целого. Задача сводится к вычислению

$$\min E\eta \tag{9}$$

по $n \in N$, после чего оптимальный размер пакета $\tilde{\upsilon}$ находится по формуле $\tilde{\upsilon} = L/\tilde{n}$, где $\tilde{n} \in N$ — решение задачи (9).

В соответствии с (2) и (8)

$$E\eta = \left(\frac{L}{n} + \tau\right) \sum_{m=n}^{\infty} \left(\frac{m}{l}\right)^{+} C_{m-1}^{n-1} p^n q^{m-n} ,$$

откуда, учитывая, что $mC_{m-1}^{n-1} = nC_m^n$,

$$E\eta = \left(\frac{L}{n} + \tau\right) \sum_{m=n}^{\infty} \left(\frac{m}{l}\right)^{+} \frac{n}{m} C_m^n p^n q^{m-n} =$$

$$= \left(\frac{1}{l}\right) (L + n\tau)\, p^n \sum_{m=n}^{\infty} \frac{(m/l)^{+}}{m/l} C_m^n q^{m-n} . \quad (10)$$

В соответствии с (2)

$$Ev = \sum_{m=n}^{\infty} mC_{m-1}^{n-1} p^n q^{m-n} = np^n \sum_{m=n}^{\infty} C_m^n q^{m-n} . \quad (11)$$

С другой стороны, случайная величина v является суммой n независимых, одинаково распределенных случайных величин, каждая из которых имеет геометрическое распределение с параметром p и математическим ожиданием $1/p$. Соответственно, $Ev = n/p$, откуда с учетом (11) следует

$$\sum_{m=n}^{\infty} C_m^n q^{m-n} = \frac{1}{p^{n+1}} . \quad (12)$$

Из (10) и (12) следует

$$E\eta = \frac{1}{l} (L + n\tau)\, p^n \sum_{m=n}^{\infty} \left[1 + \frac{(m/l)'}{m/l}\right] C_m^n q^{m-n} ,$$

где $z' = z^{+} - z$, откуда

$$E\eta = \frac{1}{l} (L + n\tau) \exp\left(\frac{\lambda L}{n}\right) (1 + \delta_l(n)) , \quad (13)$$

где

$$\delta_l(n) = \sum_{m=n}^{\infty} \alpha_{nm} \frac{(m/l)'}{m/l} , \quad (14)$$

где коэффициенты $\alpha_{nm} = p^{n+1} C_m^n q^{m-n}$, $p = e^{-\lambda L/n}$, $q = 1 - p$. При этом в соответствии с (12)

$$\sum_{m=n}^{\infty} \alpha_{nm} = 1 . \quad (15)$$

Из (14) и (15) видно, что

$$0 < \delta_l(n) < \frac{l}{n} . \quad (16)$$

Целевая функция (13) для общего случая $l \geq 1$ совпадает с целевой функцией в (3) для случая $l = 1$ с точностью до множителя $(1/l)\,[1 + \delta_l(n)]$, откуда с учетом (16) видно, что полученное выше решение для случая $l = 1$ практически дает и решение для случая $l > 1$, если оптимальное число пакетов n достаточно велико.

Обозначим через

$$f_l(n) = \frac{f(n)}{l} [1 + \delta_l(n)] \quad (17)$$

целевую функцию (13) — среднее время выполнения задания при данных значениях n — числе пакетов и l — числе элементов, где $f(n) = (L + n\tau) \exp(\lambda L/n)$ — целевая функция (6) для случая $l = 1$.

Задача выбора оптимального размера пакета $\tilde{v}$ сводится к нахождению

$$\min f_l(n) = f_l(\tilde{n}_l) . \quad (18)$$

Здесь минимум берется по всем $n \in N$, где N — множество допустимых значений n (например, N — множество целочисленных значений n, кратных 2, лежащих в некотором допустимом диапазоне, и т. п.). Полагаем $\tilde{v} = L/\tilde{n}_l$, где $\tilde{n}_l$ — решение задачи (18). Введем также дополнительную задачу нахождения

$$\min f_l(n) = f_l(n_l^*) , \quad (19)$$

где минимум берется по всем (не только целочисленным) значениям $n \geq 1$. Далее оптимальный размер пакета v_l^* (без ограничений целочисленности $n \in N$) определим по формуле $v_l^* = L/n_l^*$, где n_l^* — решение задачи (19). Из выражений (13)–(16) далее следует теорема 3.

Теорема 3. *Решение оптимизационной задачи* (19) *удовлетворяет неравенствам*

$$\frac{f_l(n^*)}{1 + \varepsilon} \leq \min_{n \geq 1} f_l(n) \leq f_l(n^*) , \quad (20)$$

где n^* — решение этой задачи для случая $l = 1$, $\varepsilon = \delta_l(n^*) < l/n^*$. Решение оптимизационной задачи (18) *удовлетворяет аналогичным неравенствам*

$$\frac{f_l(\tilde{n})}{1 + \varepsilon} \leq \min_{n \in N} f_l(n) \leq f_l(\tilde{n}) , \quad (21)$$

где $\tilde{n}$ — решение этой задачи для случая $l = 1$, $\varepsilon = \delta_l(\tilde{n}) < l/\tilde{n}$.

Доказательство. Равенство (17) при $n = n^*$ имеет вид:

$$f_l(n^*) = \frac{f(n^*)}{l} [1 + \delta_l(n^*)] . \quad (22)$$

Из этого же равенства, учитывая, что $\delta_l(n) > 0$, следует

$$f_l(n) \geq \frac{f(n)}{l},$$

откуда с учетом (22)

$$\min_{n\geq 1} f_l(n) \geq \frac{1}{l} \min_{n\geq 1} f(n) = \frac{f(n^*)}{l} = \frac{f_l(n^*)}{1+\delta_l(n^*)},$$

что вместе с (16) доказывает левое неравенство в (20). Правое неравенство очевидно. Доказательство неравенств (21) аналогично. Теорема доказана.

Таким образом, полученное решение для случая $l = 1$ практически дает решение и в случае $l > 1$, если число пакетов много больше по сравнению с количеством элементов l.

3 Заключение

Получено решение указанной выше основной проблемы (выбора оптимального размера пакета) для модели со сбоями элементов в естественной с прикладной точки зрения асимптотике, а именно для случая высоконадежных элементов и при малом времени пересылки пакетов. Существенно, что полученное решение не зависит от общего объема всего задания, что, в частности, позволяет использовать его в ситуации неопределенности, когда эта величина, вообще говоря, может быть неизвестной и случайной. Отметим также, что представляет интерес дальнейшее обобщение полученных результатов на ситуацию, когда различные элементы могут иметь существенно различные характеристики как производительности, так и надежности, а также на модель с отказами и восстановлением (заменой) отказавших элементов.

Литература

1. *Ронжин А. Ф., Суриков В. Н.* О времени полного перебора // Обозр. прикл. пром. матем., 2007. Т. 14. № 3. С. 506–508.
2. *Коновалов М. Г., Малашенко Ю. Е., Назарова И. А.* Модели и методы управления заданиями в системах распределенных вычислительных ресурсов. — М.: ВЦ РАН, 2009. 110 с. (Сообщения по прикладной математике.)
3. *Коновалов М. Г., Малашенко Ю. Е., Назарова И. А.* Оперативное управление потоком заданий в системе распределенных вычислительных ресурсов // VI Московская междунар. конф. по исследованию операций: ORM-2010: Труды. — М.: МАКС Пресс, 2010. С. 301–302.
4. *Козлов М. В., Малашенко Ю. Е., Назарова И. А., Ронжин А. Ф.* Анализ режимов управления вычислительным комплексом в условиях неопределенности. — М.: ВЦ РАН, 2011. 63 с. (Сообщения по прикладной математике.)
5. *Коновалов М. Г., Малашенко Ю. Е., Назарова И. А.* Управление заданиями в гетерогенных вычислительных системах // Известия РАН. Теория и системы управления, 2011. № 2. С. 72–90.
6. *Гнеденко Б. В., Беляев Ю. К., Соловьев А. Д.* Математические методы в теории надежности. — М.: Наука, 1965. 524 с.
7. *Gnedenko B. V., Pavlov I. V., Ushakov I. A.* Statistical reliability engineering. — N.Y.: John Wiley, 1999. 514 p.

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2012. Т. 6. Вып. 2. С. 101–112

НЕЧЕТКИЕ ПЕРЕМЕННЫЕ КАК СПОСОБ ФОРМАЛИЗАЦИИ ХАРАКТЕРИСТИК ПОГРЕШНОСТИ В ЗАДАЧАХ МАТЕМАТИЧЕСКОЙ ОБРАБОТКИ

К. К. Семенов[1]

Аннотация: В метрологии периодически высказываются предложения об использовании в измерительной практике теории нечетких переменных. Известны примеры успешного внедрения ее результатов и основных идей, которые убеждают в перспективности подобного подхода, хотя и носят частный характер. Важнейшим приложением теории нечетких множеств в метрологии является учет и рассмотрение плохо формализуемой информации об источниках неопределенности, как правило, традиционными методами не учитывающейся. В настоящей работе представлены результаты теоретического исследования принципиальной возможности другого приложения: учета и рассмотрения средствами теории нечетких переменных традиционных в метрологии хорошо формализуемых сведений о погрешности. Показана возможность единообразного описания погрешности при помощи нечетких множеств, не противоречащего сложившейся в отечественной метрологии традиции и требованиям норм и стандартов.

Ключевые слова: нечеткие переменные; характеристики погрешности; результаты измерений

1 Формулировка проблемы

Предложения об использовании теории нечетких множеств или ее частных случаев в теории и практике измерений и математической обработки появились достаточно давно [1–4], возможность использования ее средств и идей исследовалась и обсуждалась в работах [5–7].

Согласно [8], «погрешность — идеализированное понятие, и погрешность не может быть известна точно. . . неопределенность результата измерения отражает отсутствие точного знания значения измеряемой величины». Данное утверждение выражает основные предпосылки к использованию теории нечетких множеств в метрологии.

Любые измерительные задачи и задачи, связанные с вычислениями с неточно заданными исходными данными, сопряжены с неопределенностью или с нечеткостью, складывающейся из многих составляющих, обусловленных различными причинами. Традиционные методы метрологии позволяют учесть только хорошо формализуемую информацию о влияющих факторах, остальные сведения, как правило, в расчет не берутся. Появление теории нечетких множеств дало возможность также ввести в рассмотрение плохо формализуемую информацию, связанную, например, с опытом экспериментатора и позволяющую повысить точность измерений [5]. К попыткам использования теории нечетких множеств в практике измерений подталкивает необходимость согласования априорной информации об измеряемых величинах, которую зачастую принципиально невозможно формализовать в рамках теории вероятностей, с операциями по оценке характеристик погрешности результатов измерений, которые выполняются в рамках математической статистики.

На сегодняшний день нечеткая интерпретация неопределенности сведений нашла широкое применение в экспертных системах и системах нечеткого контроля и управления. Однако используемые в них механизмы работы с нечеткими переменными не позволяют решать задачи статистической или иной обработки результатов прямых измерений, поскольку плохо соотносятся с теорией вероятностей.

В измерительной практике сложилась ситуация, когда основную математическую обработку в метрологических задачах выполняют традиционными методами математической статистики, а при помощи средств и методов теории нечетких множеств реализуется обработка плохо формализуемой информации и уточнение результатов. Подобное разделение может быть преодолено, если удастся *реализовать в рамках теории нечетких переменных традиционные методы метрологии*, что и является целью настоящей статьи. Достичь этого предполагается представлением характеристик погрешности как нечетких переменных, выполненным с учетом специфики обработки результатов измерений и принятых на сегодняшний день метрологических норм.

[1]Санкт-Петербургский государственный политехнический университет, semenov.k.k@gmail.com

2 Предъявляемые требования к представлению характеристик погрешности результатов измерений нечеткими переменными

Результат x любого измерения сопровождается погрешностью Δx, в общем случае складывающейся из систематической $\Delta_{\text{сист}}x$ и случайной $\Delta_{\text{случ}}x$ составляющих. Точное значение $x_{\text{ист}}$ измеряемой физической величины получено быть не может. Следовательно, не может быть известно и фактическое значение $\Delta x = x - x_{\text{ист}} = \Delta_{\text{сист}}x + \Delta_{\text{случ}}x$ погрешности измерения. В связи с этим в метрологии используют пределы допускаемых значений для погрешности результатов измерений.

В качестве подобной характеристики для систематической составляющей погрешности используются границы интервала $J_1 = [-\Delta_{\text{сист}}, \Delta_{\text{сист}}]$, гарантированно накрывающего значение $\Delta_{\text{сист}}x$: $|\Delta_{\text{сист}}x| < \Delta_{\text{сист}}$. Для случайной же составляющей погрешности используют границы интервалов $J_P = [-\Delta_{\text{случ}}, \Delta_{\text{случ}}]$, таких что с вероятностью не менее P (обычно $0{,}8 \div 0{,}95$) случайная составляющая погрешности $\Delta_{\text{случ}}x$ примет значение, лежащее внутри J_P. Таким образом, основной характеристикой, используемой для погрешности, является интервал, границы которого служат пределами допускаемых значений при заданной доверительной вероятности, как это установлено в [9]. *Интервальное описание характеристик погрешностей должно быть сохранено при построении их представления нечеткими переменными*.

Естественным расширением понятия интервала может служить понятие нечеткого интервала, введенное в [4]. Соответствующий ему формализм позволяет оперировать с набором интервалов, каждому из которых поставлено в соответствие значение $0 \le \alpha \le 1$ меры, именуемой уровнем доверия [10]. Наблюдаемая аналогия с понятиями, принятыми в теории вероятностей, делает актуальными попытки представления в рамках теории нечетких множеств как совокупности $\langle \Delta_{\text{сист}}x, \Delta_{\text{случ}}x \rangle$ составляющих погрешности, так и самих результатов измерений в целом, поскольку сохраняют интервальный принцип их описания. Необходимо только, чтобы при этом не возникало противоречий с правилами работы с систематической и случайной составляющими погрешности, принятыми в метрологии.

На практике для операций с систематическими составляющими погрешности $\Delta_{\text{сист}}x$ используется интервальная арифметика, а для операций со случайными составляющими погрешности $\Delta_{\text{случ}}x$ используются результаты теории вероятностей. Унифицировать операции с совокупностью $\langle \Delta_{\text{сист}}x, \Delta_{\text{случ}}x \rangle$ так, чтобы в рамках одного из перечисленных аппаратов одновременно работать как с систематической составляющей погрешности, так и со случайной составляющей не представляется возможным. Теория нечетких переменных же позволяет осуществить подобную унификацию.

Поскольку обработка составляющих $\Delta_{\text{сист}}x$ и $\Delta_{\text{случ}}x$ погрешности выполняется с помощью различных правил и операций, то можно сформулировать набор требований, предъявляемых к разрабатываемому представлению погрешностей как нечетких переменных, следующим образом:

(1) в рамках подобного представления *должно осуществляться разделение систематической и случайной составляющих погрешности*, для каждой из которых обеспечиваются соответствующие правила работы с их интервальными характеристиками, в частности:

(2) *аппарат теории нечетких переменных при использовании указанного представления должен обеспечивать уменьшение среднеквадратического отклонения либо интервала неопределенности случайной составляющей погрешности в $\sqrt{n}$ раз при усреднении n результатов прямых многократных измерений значения одной и той же физической величины*;

(3) *систематическая составляющая погрешности должна обрабатываться по правилам интервальной арифметики*.

Формулировка требования 2 обусловлена тем, что зачастую при метрологическом анализе результатов математической обработки неточных данных применяют линеаризацию вычисляемых функций. По этой же причине в дальнейшем ограничимся рассмотрением только линейных операций с нечеткими переменными.

Чтобы обладать практической ценностью в задачах оценки характеристик погрешности результатов математической обработки, представление должно также удовлетворять следующим требованиям:

(4) *использование представления погрешности результатов прямых измерений как нечетких переменных не должно увеличивать время обработки данных по сравнению с традиционными методами метрологии*;

(5) *значения характеристик погрешности конечных результатов математической обработки должны совпадать или быть близки к оценкам, которые могут быть получены в рамках традиционных методов*.

Данный набор требований исчерпывающе описывает необходимый для практики набор свойств представления погрешностей как нечетких переменных. Прежде чем выполнить теоретический анализ перечисленных требований, кратко напомним основные понятия теории нечетких переменных.

3 Основные понятия теории нечетких переменных

Нечеткой переменной ξ является совокупность пар $\langle x, \mu_\xi(x)\rangle$, где $x \in R$ — возможные значения ξ, а функция $\mu_\xi(x)$, поставленная им в соответствие, принимает значения из интервала $[0, 1]$. Функцию $\mu_\xi(x)$ называют функцией принадлежности, она является основной характеристикой нечеткой переменной ξ. Носителем S_ξ нечеткой переменной часто называют интервал $[a, b]$, такой что для любого числа $x \in [a, b]$ известно, что $\mu_\xi(x) \neq 0$, а для любого числа $y \notin [a, b]$ известно, что $\mu_\xi(y) = 0$ [11]. Значения $\mu_\xi(x)$ принято называть уровнем доверия.

В работе [10] введено понятие вложенного интервала J_α функции принадлежности как интервала, границами которого являются две точки пересечения прямой $\alpha = const$, параллельной оси x, с графиком функции $\mu_\xi(x)$. Пример функции принадлежности с обозначением перечисленных ее характеристик представлен на рис. 1.

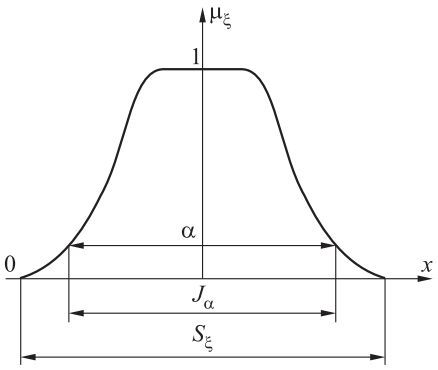


Рис. 1 Пример функции принадлежности $\mu_\xi(x)$ нечеткой переменной ξ

4 Анализ сформулированных требований

Памятуя о поставленной в работе цели, можно выполнить представление погрешности нечеткой переменной двумя путями:

(1) просто указав соответствие между основными интервальными характеристиками ($\Delta_{\text{сист}}$ и $\Delta_{\text{случ}}$) погрешности и сопоставленной ей функцией принадлежности,

(2) указав содержательную интерпретацию функции принадлежности как варианта описания погрешности в целом.

Второй вариант является более предпочтительным, поскольку помимо реализации традиционных методов метрологии позволяет достичь единства понятийного аппарата при работе с погрешностями и нечеткими переменными.

В [11] отмечается, что функция принадлежности имеет субъективный характер и может интерпретироваться на основе понятия вероятности: значение $\mu_\xi(x)$ может быть понято как вероятность того, что лицо, принимающее решение, отнесет значение x к множеству значений ξ.

Исходя из данной содержательной физической интерпретации функции принадлежности, при представлении погрешности нечеткой переменной естественным представляется трактовать значения уровня доверия α как степень уверенности в том, что интервал J_α полностью накрывается интервалом возможных значений погрешности Δx или интерквантильным промежутком заданной вероятностной меры. При такой интерпретации сохраняется интервальный принцип описания характеристик погрешности: вложенные интервалы определяют множества возможных значений полной погрешности при заданном уровне доверия.

В связи с указанной трактовкой уместно потребовать, чтобы $J_{\alpha_1} \subseteq J_{\alpha_2}$ при $\alpha_1 \geq \alpha_2$. Из этого следует, что функция принадлежности $\mu_\xi(x)$, выражающая сведения о значениях погрешности, должна принадлежать к классу нормальных функций принадлежности [12, 13], т. е. таких функций, для которых выполнены следующие условия:

– существует промежуток значений носителя $(a, b]$, $a \leq b$, во всех точках которого функция принадлежности равна единице $\mu_\xi(x_0) = 1$, а также

– при отступлении влево от точки $x = a$ или вправо от точки $x = b$ функция принадлежности $\mu_\xi(x)$ не возрастает.

Таким образом, функция принадлежности нечеткой переменной, выражающей характеристики погрешности результатов прямых измерений, должна являться в общем случае криволинейной трапецией.

4.1 Анализ требования 1

Впервые представление погрешности нечеткими переменными было выполнено Резником с соавторами в работе [4] и отражало возможное наличие систематической $\Delta_{\text{сист}}x$ и случайной $\Delta_{\text{случ}}x$ составляющих погрешности результата измерений. Вслед за указанной работой в [14] предложено считать, что вложенный интервал J_1, т. е. верхнее основание трапеции функции принадлежности, отражает сведения о $\Delta_{\text{сист}}x$. Например, если известно, что $|\Delta_{\text{сист}}x| \le \Delta_{\text{сист}}$, то J_1 целесообразно брать равным $J_1 = [-\Delta_{\text{сист}},\, \Delta_{\text{сист}}]$.

Вложенные интервалы при прочих значениях уровня доверия $\alpha \ne 1$ предлагается считать отражающими сведения о характере случайной составляющей погрешности $\Delta_{\text{случ}}x$.

Таким образом, *для любой имеющейся функции принадлежности всегда может быть осуществлено разделение выражаемых ею сведений о характеристиках систематической и случайной составляющих погрешности*. Для этого достаточно лишь рассмотреть по отдельности вложенный интервал J_1 и вложенные интервалы J_α при значениях уровня доверия $\alpha \ne 1$.

4.2 Анализ требования 2

Поскольку положение требования 2 относится только лишь к случайной составляющей погрешности, то будем считать, что рассматривается такая ситуация, когда систематическая составляющая $\Delta_{\text{сист}}x$ погрешности результата измерений x пренебрежимо мала или вовсе отсутствует, т. е. $\Delta_{\text{сист}}x = 0$.

Определим свойства функции $\mu_\xi(x)$, необходимые для анализа требования 2.

Будем считать, что математическое ожидание случайной величины, которой является случайная составляющая погрешности $\Delta_{\text{случ}}x$, есть $\mathrm{M}[\Delta_{\text{случ}}x] = 0$. В случае, если $\mathrm{M}[\Delta_{\text{случ}}x] = c \ne 0$, имеем возможность считать в качестве систематической составляющей погрешности величину $\Delta_{\text{сист}}x = \mathrm{M}[\Delta_{\text{случ}}x] = c$, а в качестве случайной составляющей — величину $(\Delta_{\text{случ}}x - c)$, математическое ожидание которой уже равно нулю.

В этом случае нечеткая переменная ξ, представляющая погрешность $\Delta x = \Delta_{\text{случ}}x$, будет иметь функцию принадлежности $\mu_\xi(x)$, равную 1, только при значении аргумента $x = 0$.

На практике чаще всего используют доверительные интервалы для значений $\Delta_{\text{случ}}(x)$ с пределами, симметричными относительно $\mathrm{M}[\Delta_{\text{случ}}x] = 0$. Поэтому будем считать, что функция $\mu_\xi(x)$ является симметричной функцией относительно значения $x = 0$ аргумента.

Так как функция $\mu_\xi(x)$ является нормальной функцией принадлежности, то при $x < 0$ она не убывает, а при $x > 0$ не возрастает.

Потребуем, чтобы функция принадлежности $\mu_\xi(x)$ являлась *аналитической функцией* на своем носителе S, что повлечет ее гладкость. Действительно, нет оснований приписывать функции принадлежности особенности, так как последняя строится экспертами исходя из физических соображений, а интуитивно наиболее подходящей на роль функции принадлежности представляется достаточно гладкая кривая. Кроме того, подобное допущение необходимо сделать в соответствии с требованием 4. Чем больше особенностей имеет функция $\mu_\xi(x)$, тем от большего числа параметров она зависит и тем большее время занимают вычисления с ней.

Переформулируем требование 2 в более удобной для дальнейшего анализа форме.

В большинстве случаев на практике в качестве пределов допускаемых значений для случайной составляющей погрешности $\Delta_{\text{случ}}x$ выбираются границы интервала $[-m\sigma,\, m\sigma]$, где σ — значение среднеквадратического отклонения $\Delta_{\text{случ}}x$. Величина m определяется на основе соображений, изложенных в [15], либо по известному неравенству П. Л. Чебышёва или его уточнениям для ряда важных практических случаев для заданного значения доверительной вероятности Q. Из теории вероятностей известно, что среднеквадратическое отклонение суммы двух независимых случайных величин с одинаковыми дисперсиями ровно в $\sqrt{2}$ раз больше среднеквадратического отклонения каждого из слагаемых.

Рассмотрим две нечеткие переменные ξ_1 и ξ_2, отражающие сведения о случайной составляющей погрешности двух результатов независимых прямых измерений значения одной и той же величины одним и тем же средством измерения. Функции принадлежности $\mu_{\xi_1}(x)$ и $\mu_{\xi_2}(x)$ указанных переменных тождественно равны, т. е. $\mu_{\xi_1}(x) = \mu_{\xi_2}(x) = \mu(x)$, носителем функции $\mu(x)$ является интервал S.

Поскольку $\mu(x)$ является такой симметричной функцией принадлежности, что при $\{x:\ x > 0,\, x \in S\}$ она не возрастает, а при $\{x:\ x \le 0,\, x \in S\}$ не убывает и при этом $\mu(x) = 1$ только при $x = 0$, то функция $\Delta_\mu(\alpha) = |\mu^{-1}(\alpha)|$ является однозначной при $1 \le \alpha < 0$. Заметим, что $\Delta_\mu(\mu(x)) = |x|$ и $\mu(\Delta_\mu(\alpha)) = \alpha$.

Обозначим $\eta = \xi_1 + \xi_2$. Так как функция принадлежности $\mu_\xi(x)$ выражает степень уверенности в том, что вложенные интервалы $J_\alpha[\xi_1]$ и $J_\alpha[\xi_2]$ накрываются промежутком $[-m\sigma,\, m\sigma]$, то, соответ-

ственно, получаем, что функция принадлежности $\mu_\eta(x)$ выражает степень уверенности в том, что вложенный интервал $J_\alpha[\eta]$ накрывается промежутком $[-m\sqrt{2}\,\sigma,\, m\sqrt{2}\,\sigma] = \sqrt{2}[-m\sigma,\, m\sigma]$. Следовательно, требование 2 может быть переформулировано в виде соотношения $\Delta_{\mu_\eta}(\alpha) = \sqrt{2}\,\Delta_\mu(\alpha)$, т. е. вложенные интервалы функции принадлежности результата суммирования должны быть в $\sqrt{2}$ раз шире, чем вложенные интервалы функции принадлежности операндов при том же самом значении уровня доверия.

Подставим в это соотношение значение $\alpha = \mu_\eta(x)$. Получим:

$$\Delta_{\mu_\eta}(\mu_\eta(x)) = \sqrt{2}\,\Delta_\mu(\mu_\eta(x))\,;$$
$$\frac{|x|}{\sqrt{2}} = \Delta_\mu(\mu_\eta(x))\,;$$
$$\mu\left(\frac{|x|}{\sqrt{2}}\right) = \mu(\Delta_\mu(\mu_\eta(x)))\,;$$
$$\mu\left(\frac{x}{\sqrt{2}}\right) = \mu_\eta(x)\,.$$

Отсюда следует, что эквивалентным требованию 2 является соблюдение соотношения:

$$\mu_\eta(x) = \mu\left(\frac{x}{\sqrt{2}}\right).$$

Проанализируем положения требования 2.

Воспользовавшись принципом обобщения Заде [11], представляющим общий вид операций над нечеткими переменными, получим соотношение, связывающее $\mu_\eta(x)$ и $\mu(x)$ и задающее правило сложения нечетких переменных:

$$\mu_\eta(z) = \sup_{\substack{x\in S,\ y\in S,\\ x+y=z}} \{\mu(x) \cap \mu(y)\}\,.$$

В силу непрерывности и монотонности функции $\mu(x)$ на интервалах $\{x:\ x > 0, x \in S\}$ и $\{x:\ x \le 0, x \in S\}$

$$\sup_{\substack{x\in S,y\in S,\\ x+y=z}} \{\mu(x) \cap \mu(y)\} = \max_{\substack{x\in S,y\in S,\\ x+y=z}} \{\mu(x) \cap \mu(y)\}\,.$$

Таким образом,

$$\max_{\substack{x\in S,y\in S,\\ x+y=z}} \{\mu(x) \cap \mu(y)\} = \mu\left(\frac{z}{\sqrt{2}}\right). \qquad (1)$$

Среди наиболее типовых вариантов введения операции пересечения и, как следствие, правил сложения нечетких переменных, встречающихся на практике, следующие [13]:

– минимаксная операция

$$\mu_{\xi_1+\xi_2}(z) = \sup_{\substack{x\in S_1,y\in S_2,\\ x+y=z}} \{\min\{\mu_{\xi_1}(x), \mu_{\xi_2}(y)\}\}\,;$$

– алгебраическая операция

$$\mu_{\xi_1+\xi_2}(z) = \sup_{\substack{x\in S,y\in S,\\ x+y=z}} \{\mu_{\xi_1}(x)\mu_{\xi_2}(y)\}\,.$$

Проверим, удовлетворяют ли перечисленные операции требованию (1).

Результату минимаксной операции сложения двух нечетких переменных с одинаковыми функциями принадлежности будет соответствовать функция принадлежности

$$\mu_\eta(z) = \sup_{\substack{x\in S,y\in S,\\ x+y=z}} \{\min\{\mu(x), \mu(y)\}\} =$$
$$= \max_{x\in S}\{\min\{\mu(x), \mu(z-x)\}\} =$$
$$= \max_{x\in S}\{\min\{\mu(x), \mu(x-z)\}\} = \mu\left(\frac{z}{2}\right).$$

Действительно, при $x_0 = z/2$ имеем $\mu(x_0) = \mu(x_0 - z)$ и, соответственно,

$$\min\{\mu(x_0), \mu(x_0 - z)\} = \mu\left(\frac{z}{2}\right).$$

При $|x| \le |x_0|$ имеем $|x - z| \ge |z|/2$ и, как следствие,

$$\mu(z-x) = \mu(|z-x|) \le \mu\left(\frac{|z|}{2}\right) = \mu\left(\frac{z}{2}\right).$$

Соответственно, получаем, что $\min\{\mu(x), \mu(x - -z)\} \le \mu(z/2)$. При $|x| \ge |x_0|\,\mu(x) \le \mu(x_0) = \mu(z/2)$ и снова $\min\{\mu(x), \mu(x-z)\} \le \mu(z/2)$. Таким образом, минимаксная операция не может удовлетворить требованию 2, поскольку $\mu(z/2) \ge \mu(z/\sqrt{2})$, причем равенство достигается только при $z = 0$.

Алгебраическое же правило

$$\mu_\eta(z) = \sup_{\substack{x\in S,y\in S,\\ x+y=z}} \{\mu_{\xi_1}(x)\mu_{\xi_2}(y)\} =$$
$$= \max_{\substack{x\in S,y\in S,\\ x+y=z}} \{\mu_{\xi_1}(x)\mu_{\xi_2}(y)\}$$

оказывается правилом определения операции сложения нечетких переменных, потенциально позволяющим удовлетворить требованию 2.

Отдельно отметим, что согласно [11] именно алгебраический вариант принципа обобщения Заде соответствует вероятностной трактовке функции принадлежности и, таким образом, наиболее близок к поставленным в настоящей работе задачам.

Может быть сформулирована и доказана следующая

Теорема. *Пусть функция $\mu(x)$ является функцией принадлежности нечеткой переменной, выражающей*

сведения о случайной погрешности результата измерений, и выбрано алгебраическое правило сложения нечетких переменных. Требование 2 *выполняется тогда и только тогда, когда*

$$\mu(x) = \exp\left[-\frac{x^2}{2\sigma^2}\right],$$

где $\sigma > 0$.

Перед тем как привести доказательство представленного утверждения, отметим, что имеет место

Лемма. *Если функция принадлежности* $\mu(x)$ *нечеткой переменной, выражающей сведения о случайной погрешности результата измерений, такова, что* $\mu(x/\sqrt{2}) = \mu^2(x/2)$ *в некоторой окрестности точки* $x = 0$*, то на всем носителе* S *выполнено* $\mu(x) = \exp[-x^2/(2\sigma^2)]$*, где* $\sigma > 0$.

Доказательство леммы приведено в приложении к статье.

Доказательство теоремы. *Необходимость.* Покажем, что из условия $\max\limits_{\substack{x\in S, y\in S,\\ x+y=z}}\{\mu(x)\mu(y)\} = \mu(z/\sqrt{2})$ следует, что $\mu(x) = e^{\alpha x^2/2}$. Введем функцию $h(x,y) = \mu(x)\mu(y)$. Наложим условие $x + y = z$. Рассмотрим задачу поиска условного максимума $\varphi(z) = \max\limits_{\substack{x\in S, y\in S,\\ x+y=z}}\{\mu(x)\mu(y)\} = \max\limits_{x\in S}\{\mu(x)\mu(z-x)\} = \max\limits_{x\in S}\{h(x, z-x)\}$.

Экстремумы функции $h(x, z-x)$ при заданном z достигаются при значениях аргумента $x = x_0 \in S$, при которых $h'(x_0, z - x_0) = 0$.

Так как $h'(x, z-x) = \mu'(x)\mu(z-x) - \mu(x)\mu'(z-x)$, то

$$\mu'(x_0)\mu(z-x_0) - \mu(x_0)\mu'(z-x_0) = 0;$$
$$\mu'(x_0)\mu(z-x_0) = \mu(x_0)\mu'(z-x_0);$$
$$\frac{\mu'(x_0)}{\mu(x_0)} = \frac{\mu'(z-x_0)}{\mu(z-x_0)}. \qquad (2)$$

Очевидно, что равенство (2) выполняется при $x_0 = z/2$, т. е. в точке с указанной абсциссой достигается экстремум функции $h(x,\ z-x)$.

В силу четности функции $\mu(x)$ выполняется равенство $\mu(x) = \mu(-x)$. Соответственно, функция $\mu'(x)$ является нечетной и $\mu'(x) = -\mu'(-x)$. Следовательно,

$$\frac{\mu'(z-x_0)}{\mu(z-x_0)} = -\frac{\mu'(x_0-z)}{\mu(x_0-z)}$$

и равенство (2) приобретает вид:

$$\frac{\mu'(x_0)}{\mu(x_0)} = -\frac{\mu'(x_0-z)}{\mu(x_0-z)}.$$

Функция $\mu'(x)/\mu(x) > 0$ при $\{x : x < 0, x \in S\}$ и $\mu'(x)/\mu(x) < 0$ при $\{x :\ x > 0, x \in S\}$. Отсюда следует, что при $z = 0$ кривые $\mu'(x)/\mu(x)$ и $-\mu'(x)/\mu(x)$ пересекаются только в начале координат. Также отсюда следует, что существует такой интервал $[-a, a] \subset S$, $a > 0$, на котором функция $\mu'(x)/\mu(x)$ является монотонно убывающей. Значит, функция $-\mu'(x)/\mu(x)$ на этом интервале будет монотонно возрастать и при $|z| \le a$ уравнение $\mu'(x_0)/\mu(x_0) = -\mu'(x_0-z)/\mu(x_0-z)$ гарантированно имеет единственное решение. Как было отмечено выше, им является $x_0 = z/2$.

Таким образом, при $|z| \le a$ имеем $\mu_\eta(z) = \max\limits_{\substack{x\in S, y\in S,\\ x+y=z}}\{\mu(x)\mu(y)\} = \mu(x_0)\mu(z-x_0) = \mu^2(z/2)$. Так как по условию теоремы $\mu_\eta(z) = \mu(z/\sqrt{2})$, то $\mu^2(z/2) = \mu(z/\sqrt{2})$. В соответствии с леммой, упомянутой выше, из этого следует, что аналитическая функция $\mu(x)$ совпадает с нормированной гауссианой $\exp\left[-x^2/(2\sigma^2)\right]$ при некотором значении параметра σ во всех точках носителя S.

Необходимость доказана.

Достаточность. Покажем, что если $\mu(x) = e^{-x^2/(2\sigma^2)}$, то $\max\limits_{\substack{x\in S, y\in S,\\ x+y=z}}\{\mu(x)\mu(y)\} = \mu(z/\sqrt{2})$.

Выполним поиск максимума выражения

$$\max_{\substack{x\in S, y\in S,\\ x+y=z}}\{\mu(x)\mu(y)\} = \max_{\substack{x\in S, y\in S,\\ x+y=z}}\left\{e^{-x^2/(2\sigma^2)}e^{-(z-x)^2/(2\sigma^2)}\right\}.$$

Искомый максимум достигается в точке максимума функции $f(x) = -x^2/(2\sigma^2) - (z-x)^2/(2\sigma^2)$. Тогда

$$\left.\frac{df(x)}{dx}\right|_{x=x_0} = -\frac{x_0}{\sigma^2} + \frac{z-x_0}{\sigma^2} = 0;$$
$$x_0 = \frac{z}{2}.$$

Таким образом,

$$\max_{\substack{x\in S, y\in S,\\ x+y=z}}\{\mu(x)\mu(y)\} = \mu^2\left(\frac{z}{2}\right).$$

С другой стороны,

$$\mu^2\left(\frac{z}{2}\right) = e^{-(z/2)^2/(2\sigma^2)}e^{-(z/2)^2/(2\sigma^2)} =$$
$$= e^{-(z/2)^2/\sigma^2} = e^{-(z/\sqrt{2})^2/(2\sigma^2)} = \mu\left(\frac{z}{\sqrt{2}}\right).$$

Следовательно,

$$\max_{\substack{x\in S,y\in S,\\ x+y=z}}\{\mu(x)\mu(y)\}=\mu\left(\frac{z}{\sqrt{2}}\right).$$

Достаточность доказана.

Таким образом, для того чтобы предъявленные к разрабатываемому представлению требования были выполнены, необходимо, чтобы функции принадлежности нечетких переменных, выражающих характеристики погрешностей результатов измерений, имели вид криволинейной трапеции, боковые стороны которой являются левой и правой половинами нормированной гауссианы.

В качестве следствия заметим, что результаты операций сложения и вычитания нечетких переменных по правилу

$$\mu_{\xi_1+\xi_2}(z)=\sup_{\substack{x\in S,y\in S,\\ x+y=z}}\{\mu_{\xi_1}(x)\mu_{\xi_2}(y)\}$$

в случае, когда операнды имеют функциями принадлежности нормированные гауссианы, также являются нормированными гауссианами, *что обеспечивает унификацию операции сложения для случайной составляющей погрешности*.

Заметим, что представленные результаты могут быть обобщены на случай, когда функция принадлежности случайной составляющей погрешности результата прямого измерения не является симметричной функцией.

4.3 Анализ требования 3

Пусть ξ_1 и ξ_2 — две нечеткие переменные с функциями принадлежности соответственно:

$$\mu_{\xi_1}(x)=\begin{cases}1, & x\in[a_1,b_1];\\ 0, & x\notin[a_1,b_1];\end{cases}$$

$$\mu_{\xi_2}(x)=\begin{cases}1, & x\in[a_2,b_2];\\ 0, & x\notin[a_2,b_2].\end{cases}$$

Переменные ξ_1 и ξ_2 представляют собой систематические погрешности $\Delta_{\text{сист}}x_1$ и $\Delta_{\text{сист}}x_2$ результатов измерений x_1 и x_2.

По итогам анализа требования 2 сделан вывод о том, что правилом, задающим функцию принадлежности суммы двух нечетких переменных, должно являться алгебраическое правило.

Пусть $\eta=\xi_1+\xi_2$. Тогда функция принадлежности нечеткой переменной η есть

$$\mu_\eta(z)=\sup_{\substack{x\in S,y\in S,\\ x+y=z}}\{\mu_{\xi_1}(x)\mu_{\xi_2}(y)\}=$$
$$=\begin{cases}1, & x\in[a_1+a_2,\,b_1+b_2];\\ 0, & x\notin[a_1+a_2,\,b_2+b_2].\end{cases}$$

Границы вложенного интервала $J_1[\eta]$ в точности совпадают с теми, которые могут быть получены при помощи интервальной арифметики. *Таким образом, алгебраическое правило сохраняет форму функций принадлежности нечетких переменных, выражающих систематическую составляющую погрешности, и позволяет унифицированно их обрабатывать.*

Тот же вывод следует из предложенной интерпретации функции принадлежности погрешности, выраженной как нечеткая переменная. Поскольку $J_1[\xi_1]=[a_1,b_1]$ и $J_1[\xi_2]=[a_2,b_2]$ интерпретируются как такие интервалы, про которые известно, что они лежат внутри интервалов возможных значений выражаемых соответственно ξ_1 и ξ_2 погрешностей, то и про интервал $J_1[\xi_1+\xi_2]=[a_1+a_2,b_1+b_2]$ известно, что он лежит внутри интервала возможных значений величины $\xi_1+\xi_2$, что полностью согласуется с требованиями интервальной арифметики.

4.4 Анализ требований 4 и 5

Пусть $\xi_{\text{сист}}$ — нечеткая переменная, функция принадлежности $\mu_{\xi_{\text{сист}}}(x)$ которой равна 1 в любой точке интервала $[-\Delta_{\text{сист}},\Delta_{\text{сист}}]$ и нулю во всех точках вне данного отрезка. Пусть $\xi_{\text{случ}}$ — нечеткая переменная, функция принадлежности $\mu_{\xi_{\text{случ}}}(x)$ которой является нормированной гауссианой, т. е. $\mu_{\xi_{\text{случ}}}(x)=\exp\left[-x^2/(2\sigma^2)\right]$. Тогда, очевидно, их сумма $\xi=\xi_{\text{сист}}+\xi_{\text{случ}}$ имеет функцию принадлежности

$$\mu_\xi(z)=\max_{\substack{x\in S_{\xi_{\text{сист}}},y\in S_{\xi_{\text{случ}}},\\ x+y=z}}\{\mu_{\xi_{\text{сист}}}(x)\mu_{\xi_{\text{случ}}}(y)\}=$$
$$=\begin{cases}\mu_{\xi_{\text{случ}}}(z+\Delta_{\text{сист}})=\exp\left[-\dfrac{(z+\Delta_{\text{сист}})^2}{2\sigma^2}\right] & \\ \qquad\qquad\text{при } z\le-\Delta_{\text{сист}};\\ \mu_{\xi_{\text{сист}}}(z)=1 \qquad \text{при } -\Delta_{\text{сист}}\le z\le\Delta_{\text{сист}};\\ \mu_{\xi_{\text{случ}}}(z-\Delta_{\text{сист}})=\exp\left[-\dfrac{(z-\Delta_{\text{сист}})^2}{2\sigma^2}\right] & \\ \qquad\qquad\text{при } z\ge\Delta_{\text{сист}}.\end{cases}$$

Учитывая, что было решено считать вложенные интервалы J_1 на уровне значимости $\alpha=1$ отражающими сведения о систематической составляющей $\Delta_{\text{сист}}x$ погрешности результата измерений, а функции принадлежности как у $\xi_{\text{случ}}$ — отражающими сведения о случайной составляющей $\Delta_{\text{случ}}x$, то получаем, что нечеткая переменная $\xi=\xi_{\text{сист}}+\xi_{\text{случ}}$

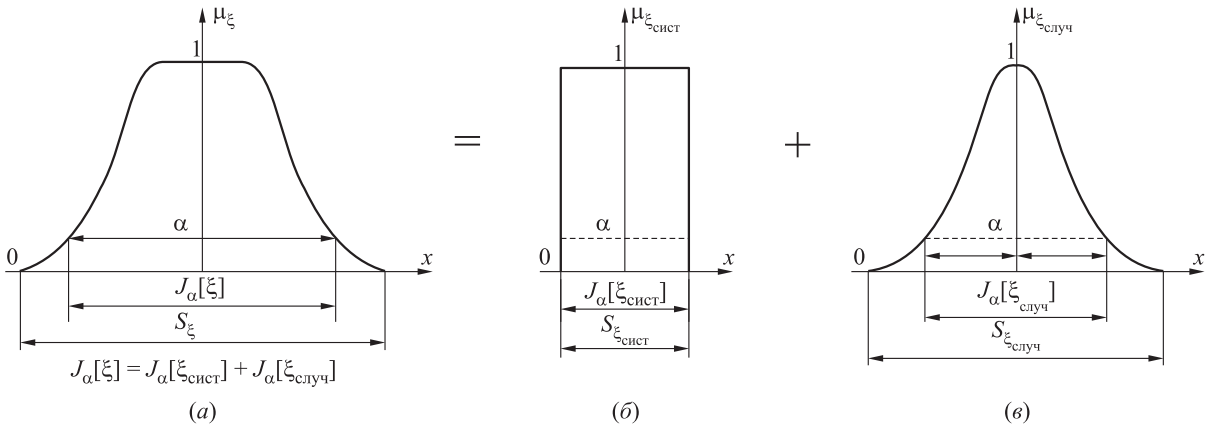


Рис. 2 Разложение нечеткой переменной, выражающей характеристики полной погрешности результата измерений (*а*), в сумму нечетких переменных, описывающих систематическую (*б*) и случайную (*в*) составляющие погрешности

отражает сведения о погрешности в целом. Заметим, что при подобном представлении случайная и систематическая составляющие погрешности всегда разделяются, как это было отмечено при анализе требования 1. Данное свойство иллюстрирует рис. 2.

Таким образом, функцией принадлежности нечеткой переменной, отражающей сведения о погрешности результата измерений, в общем случае является функция, график которой представлен на рис. 1 и 2, *а*.

Заметим, что в общем случае функция вида $\mu_\xi(z)$ может быть полностью описана при помощи только трех параметров: границ интервала $J_1 = [a,\, b]$, во всех точках которого она принимает значения, равные 1, и параметром масштаба σ, описывающим боковые гауссианы. Подобная естественная параметризация функции принадлежности $\mu_\xi(z)$ позволяет заменить алгебраическое правило сложения нечетких переменных, наивная реализация которого при вычислениях потребует выполнения достаточно большого числа операций, на набор простых правил для параметров.

Действительно, пусть заданы две нечеткие переменные ξ_1 и ξ_2, представленные соответственно кортежами чисел $\langle [a_1, b_1],\, \sigma_1\rangle$ и $\langle [a_2, b_2],\, \sigma_2\rangle$ и выражающие соответственно погрешности Δx_1 и Δx_2 величин x_1 и x_2. Тогда нечеткая переменная $\xi_3 = \xi_1 + \xi_2$ будет характеризоваться кортежем $\langle [a_3, b_3],\, \sigma_3\rangle$, таким что

$$
\begin{aligned}
a_3 &= a_1 + a_2\,;\\
b_3 &= b_1 + b_2\,;\\
\sigma_3 &= \sqrt{\sigma_1^2 + \sigma_2^2}\,,
\end{aligned}
$$

в чем несложно убедиться, воспользовавшись правилом $\mu_{\xi_1+\xi_2}(z) = \sup\limits_{\substack{x\in S, y\in S,\\ x+y=z}} \{\mu_{\xi_1}(x)\mu_{\xi_2}(y)\}$.

Заметим, что представленные соотношения для параметров в точности совпадают с правилом вычисления границ $[a_3,\, b_3]$ интервала допускаемых значений для суммы $\Delta_{\text{сист}}x_1 + \Delta_{\text{сист}}x_2$ систематических составляющих погрешности величин x_1 и x_2 и правилом вычисления среднеквадратического значения σ_3 суммы $\Delta_{\text{случ}}x_1 + \Delta_{\text{случ}}x_2$ случайных составляющих погрешности. Таким образом, *данные правила в точности реализуют общепринятую методику обработки характеристик погрешности результатов измерений*. Поскольку вычисления предлагается выполнять напрямую с параметрами $\langle [a, b],\, \sigma\rangle$, то *использование аппарата нечетких переменных не повлечет увеличения времени, затрачиваемого на математическую обработку результатов измерений*.

Для случая масштабирования нечеткой переменной, т. е. для $\xi_3 = c\xi_1$, где c — некоторое действительное число, отличное от нуля, результат ξ_3 будет характеризоваться кортежем $\langle [a_3, b_3],\, \sigma_3\rangle$, где

$$
\begin{aligned}
a_3 &= \min\left(ca_1,\, cb_1\right)\,;\\
b_3 &= \max\left(ca_1,\, cb_1\right)\,;\\
\sigma_3 &= |c|\sigma_1\,.
\end{aligned}
$$

Действительно, масштабирование нечеткой переменной, по сути, есть масштабирование ее носителя, т. е. ξ_3 будет соответствовать функция принадлежности $\mu_{\xi_1}(cx)$. Представленные правила также в точности повторяют соотношения, принятые в метрологической практике.

Рассмотренного набора линейных операций достаточно для применения разработанного представления

при решении такой задачи, как автоматический контроль погрешности результатов вычислений при математической обработке результатов измерений на основе линеаризации вычисляемой функции [16].

Таким образом, достигается заявленная в начале работы цель.

5 Выводы

Подытоживая, заметим, что в случае, когда требуется выполнить только линейные преобразования с результатами измерений и характеристиками их погрешностей, представление погрешности как нечетких переменных не затрудняет вычислений. Представленные в предыдущем разделе правила работы с параметрами $\langle [a,b],\, \sigma \rangle$ повторяют правила для систематической составляющей погрешности, принятые в интервальной арифметике, и правила работы с дисперсией и среднеквадратическими отклонениями, принятые в теории вероятностей для независимых случайных величин. Сохранение вида функции принадлежности для результатов линейных операций позволяет унифицировать представление характеристик погрешности нечеткими переменными. *Таким образом, представление погрешности как нечеткой переменной согласуется с* [8] *и отечественной метрологической нормативной базой*.

ПРИЛОЖЕНИЕ

Доказательство леммы

Условие леммы позволяет указать следующие свойства функции $\mu(x)$.

Так как $\mu(x)$ является функцией принадлежности нечеткой переменной, выражающей сведения о случайной погрешности результата измерений, следовательно, она является четной относительно прямой $x = 0$ функцией и при этом $\mu(x) = 1$ только при $x = 0$. Аналитическая функция $\mu(x)$ не имеет разрывов и является достаточно гладкой во всех точках своего носителя S. В точке $x = 0$ достигается ее единственный экстремум.

Из равенства $\mu^2(k_1 x) = \mu(k_2 x)$, $k_1 > 0$, $k_2 > 0$, следует, что функция $\mu(x)$ либо тождественно равна 1, либо имеет бесконечное число ненулевых слагаемых в ее разложении в ряд Маклорена. Иными словами, $\exists \mu^{(i)}(0)$ $\forall i \in N$ и при этом $\forall i \in N : \mu^{(i)}(0) = 0$ $\exists j \in N$, $j > i$: $\mu^{(i)}(0) \neq 0$.

Действительно, пусть

$$\mu(x) = \sum_{n=0}^{N} \frac{\mu^{(n)}(0)}{n!}\, x^n ,$$

где N — некоторое натуральное число либо ноль и $\mu^{(N)}(0) \neq 0$. Тогда

$$\mu^2(k_1 x) = \sum_{n=0}^{2N} a_n k_1^n x^n ,$$

причем $a_{2N} \neq 0$. В то же время

$$\mu(k_2 x) = \sum_{n=0}^{N} \frac{k_2^n \mu^{(n)}(0)}{n!}\, x^n .$$

Из условия леммы следует, что функция $\mu(x)$ константой в области своего носителя S не является; следовательно, $N > 0$. Таким образом, функция $\mu^2(k_1 x)$ является полиномом степени $2N$, а функция $\mu(k_2 x)$ — полиномом степени N. Но поскольку $\mu^2(k_1 x) = \mu(k_2 x)$ и $N > 0$, получаем противоречие. Следовательно, $N = \infty$.

Поскольку функция $\mu(x)$ является четной относительно значения $x = 0$ своего аргумента, $\mu^{(n)}(0) \neq 0$ только если n — натуральное четное число, т. е. в разложении $\mu(x)$ в ряд Маклорена будут ненулевыми только слагаемые при четных индексах:

$$\mu(x) = \sum_{n=0}^{\infty} \frac{\mu^{(2n)}(0)}{(2n)!}\, x^{2n} .$$

1. Согласно теореме Лейбница о значении высшей производной произведения достаточно гладких функций получаем:

   $$\left.\frac{d^n(\mu^2(k_1 x))}{dx^n}\right|_{x=0} = k_1^n \sum_{m=0}^{m} C_n^m \mu^{(m)}(0)\mu^{(n-m)}(0) .$$

   С другой стороны, в силу равенства $\mu^2(k_1 x) = \mu(k_2 x)$ имеем:

   $$\left.\frac{d^n(\mu^2(k_1 x))}{dx^n}\right|_{x=0} = \left.\frac{d^n(\mu(k_2 x))}{dx^n}\right|_{x=0} = k_2^n \mu^{(n)}(0) ,$$

   откуда следует, что

   $$\mu^{(n)}(0) = k^n \sum_{m=0}^{n} C_n^m \mu^{(m)}(0)\mu^{(n-m)}(0) , \qquad (\text{П.1})$$

   где

   $$k = \frac{k_1}{k_2} = \frac{1}{\sqrt{2}} .$$

   Видно, что значения всех производных $d^n\mu(x)/dx^n$ четных порядков n в точке $x = 0$ связаны со значением $\mu''(0)$ простыми алгебраическими соотношениями, например $\mu^{(4)}(0) = 3\mu''^2(0)$. Отсюда также следует, что $\mu'' \neq 0$, в противном случае $\mu(x) \equiv$ $\equiv 0$, чего быть не может по условиям доказываемого утверждения.

2. Рассмотрим функцию $h(x) = \mu'(x)/\mu(x)$ при $x \in$ $\in S$. Она является нечетной функцией относительно значения $x = 0$ своего аргумента, и, следовательно, $h^{(2n)}(0) = 0$ при всех натуральных n, т. е. только производные нечетных порядков функции $h(x)$ имеют ненулевое значение при $x = 0$.

   Обозначим через $\{a_n\}_{n=0}^{n=\infty}$ коэффициенты в разложении в ряд Маклорена функции $\mu'(x)$, через $\{b_n\}_{n=0}^{n=\infty}$ — коэффициенты в разложении в ряд Маклорена функции $\mu(x)$, а через $\{q_n\}_{n=0}^{n=\infty}$ — коэффициенты в разложении в ряд Маклорена функции $h(x)$. Как было отмечено выше, $a_{2i} = 0$ и

$b_{2i+1} = 0$ при всех $i \in N \cup \{0\}$. Таким образом, коэффициенты b_n образуют последовательность $\{1, 0, \mu^{(2)}(0)/2!, 0, \mu^{(4)}(0)/4!, 0, \ldots\}$, а коэффициенты a_n, соответственно, образуют последовательность $\{0, \mu^{(2)}(0)/1!, 0, \mu^{(4)}(0)/3!, 0, \ldots\}$, получаемую из почленного дифференцирования ряда Маклорена для функции $\mu(x)$.

Покажем методом математической индукции, что все производные нечетных порядков при натуральных $n > 1$, т. е. все производные функции $h(x)$, начиная с третьей, равны нулю в точке $x = 0$. Для этого достаточно показать, что все коэффициенты $q_n = 0$ при $n \geq 3$.

Воспользуемся обращением формулы коэффициентов произведения степенных рядов для получения значений коэффициентов q_n:

q_0: $a_0 = q_0 b_0$, откуда получаем $q_0 = a_0/b_0 = 0$;

q_1: $a_1 = q_0 b_1 + q_1 b_0$, откуда получаем $q_1 = a_1/b_0 = \mu''(0)$;

q_2: $a_2 = q_0 b_2 + q_1 b_1 + q_2 b_0$, откуда получаем $q_2 = (a_2 - q_1 b_1)/b_0 = 0$;

q_3: $a_3 = q_0 b_3 + q_1 b_2 + q_2 b_1 + q_3 b_0$, откуда получаем

$$\frac{\mu^{(4)}(0)}{6} = \frac{\mu''(0)}{2}\mu''(0) + q_3\,,\ q_3 = \frac{\mu^{(4)}(0)}{6} - \frac{\mu''^{\,2}(0)}{2}\,.$$

Но поскольку $\mu^{(4)}(0) = 3\mu''^{\,2}(0)$, как отмечалось в п. 1 настоящего доказательства, $q_3 = 0$.

База индукции: показано, что $h^{(3)}(0) = 0$.

Индукционный переход: предположим, что доказываемое утверждение верно для нечетного $n = 2r + 1$, где $r \geq 2$ — натуральное число, т. е. выполнено $q_n = 0$ или, что то же, $h^{(n)}(0) = 0$. Рассмотрим утверждение для $n + 2$:

q_n: $q_n = 0$;

q_{n+1}: $q_{n+1} = 0$;

q_{n+2}: $a_{n+2} = \sum\limits_{m=0}^{n+2} q_m b_{n-m+2} = q_1 b_{n+1} + q_{n+1} b_0$, откуда получаем

$$q_{n+2} = \frac{a_{n+2} - q_1 b_{n+1}}{b_0} = a_{n+2} - \mu''(0) b_{n+1} = (n+3) b_{n+3} - \mu''(0) b_{n+1} = \frac{\mu^{(n+3)}(0)}{(n+2)!} - \mu''(0)\frac{\mu^{(n+1)}(0)}{(n+1)!}\,.$$

Так как $q_n = 0$, то $a_n = \sum\limits_{m=0}^{n} q_m b_{n-m} = q_1 b_{n-1}$, откуда

$$\frac{\mu^{(n+1)}(0)}{n!} = \mu''(0)\frac{\mu^{(n-1)}(0)}{(n-1)!}\,.$$

Следовательно, $\mu^{(n-1)}(0) = \mu''(0)\mu^{(n-1)}(0)n$. Продолжая по аналогии, получим, что

$$\mu^{(n+1)}(0) = \mu''^{(n+1)/2}(0) \prod_{j=0}^{(n+1)/2-1} (2j+1)\,. \quad \text{(П.2)}$$

Рассмотрим полученное в п. 1 настоящего доказательства соотношение (П.1) для показателя порядка $n + 3$:

$$\mu^{(n+3)}(0) = \frac{1}{(\sqrt{2})^{n+3}} \sum_{m=0}^{n+3} C_{n+3}^m \mu^{(m)}(0)\mu^{(n-m+3)}(0)\,.$$

Отделим от суммы, стоящей в правой части равенства, первое и последнее слагаемое:

$$\mu^{(n+3)}(0) = \frac{1}{2^{(n+1)/2+1}} \sum_{m=1}^{n+2} C_{n+3}^m \mu^{(m)}(0)\mu^{(n-m+3)}(0) + 2\left(\frac{1}{2^{(n+1)/2+1}}\,\mu^{(n+3)}(0)\right);$$

$$\left(1 - \frac{1}{2^{(n+1)/2}}\right)\mu^{(n+3)}(0) = \frac{1}{2^{(n+1)/2+1}} \sum_{m=1}^{n+2} C_{n+3}^m \mu^{(m)}(0)\mu^{(n-m+3)}(0)\,;$$

$$\left(2^{(n+1)/2} - 1\right)\mu^{(n+3)}(0) = \frac{1}{2} \sum_{m=1}^{n+2} C_{n+3}^m \mu^{(m)}(0)\mu^{(n-m+3)}(0)\,.$$

Поскольку $\mu^{(k)}(0) = 0$ при всяком нечетном k, удалим из суммы нулевые слагаемые:

$$\left(2^{(n+1)/2} - 1\right)\mu^{(n+3)}(0) = \frac{1}{2} \sum_{m=1}^{(n+1)/2} C_{n+3}^{2m} \mu^{(2m)}(0)\mu^{(n-2m+3)}(0)\,.$$

Подставим (П.2) в полученное соотношение:

$$\left(2^{(n+1)/2} - 1\right)\mu^{(n+3)}(0) = \frac{1}{2}\mu''^{(n+1)/2+1}(0) \times \times \sum_{m=1}^{(n+1)/2} C_{n+3}^{2m} \prod_{j=0}^{m-1}(2j+1) \prod_{s=0}^{(n+1)/2-m}(2s+1)\,.$$

Поскольку

$$C_{n+3}^{2m} \prod_{j=0}^{m-1}(2j+1) \prod_{s=0}^{(n+1)/2-m}(2s+1) = (n+3)!\frac{\prod\limits_{j=0}^{m-1}(2j+1)}{(2m)!}\,\frac{\prod\limits_{s=0}^{(n+1)/2-m}(2s+1)}{(n-2m+3)!} = (n+3)!\frac{1}{2^m m!}\,\frac{1}{2^{(n+1)/2-m+1}((n+1)/2-m+1)!} = \frac{(n+3)!}{2^{(n+1)/2+1} m!((n+1)/2-m+1)!}\,,$$

получим:

$$\left(2^{(n+1)/2}-1\right)\mu^{(n+3)}(0)=\frac{1}{2^{(n+1)/2+2}}\times$$

$$\times\,\mu''^{(n+1)/2+1}(0)\sum_{m=1}^{(n+1)/2}\frac{(n+3)!}{m!((n+1)/2-m+1)!};$$

$$\left(2^{(n+1)/2}-1\right)\mu^{(n+3)}(0)=$$

$$=\frac{1}{2^{(n+1)/2+2}}\,\mu''^{(n+1)/2+1}(0)\frac{(n+3)!}{((n+1)/2+1)!}\times$$

$$\times\sum_{m=1}^{(n+1)/2}\frac{((n+1)/2+1)!}{m!((n+1)/2-m+1)!};$$

$$\left(2^{(n+1)/2}-1\right)\mu^{(n+3)}(0)=$$

$$=\frac{1}{2^{(n+1)/2+2}}\,\mu''^{(n+1)/2+1}(0)\frac{(n+3)!}{((n+1)/2+1)!}\times$$

$$\times\sum_{m=1}^{(n+1)/2}C^m_{(n+1)/2+1}.$$

Так как согласно биному Ньютона

$$(1+1)^n-1^n-1^n=\sum_{m=1}^{n-1}C^m_n=2^n-2$$

и так как

$$\frac{(n+3)!}{((n+1)/2+1)!}=2^{(n+1)/2+1}\prod_{j=0}^{(n+1)/2}(2j+1),$$

получим

$$\left(2^{(n+1)/2}-1\right)\mu^{(n+3)}(0)=$$

$$=\frac{1}{2^{(n+1)/2+2}}\,\mu''^{(n+1)/2+1}(0)\left(2^{(n+1)/2+1}-2\right)\times$$

$$\times\,2^{(n+1)/2+1}\prod_{j=0}^{(n+1)/2}(2j+1);$$

$$\mu^{(n+3)}(0)=\mu''^{(n+1)/2+1}(0)\prod_{j=0}^{(n+2)/2}(2j+1).$$

Значит,

$$q_{n+2}=\frac{\mu^{(n+3)}(0)}{(n+2)!}-\mu''(0)\frac{\mu^{(n+1)}(0)}{(n+1)!}=$$

$$=\frac{\mu''^{(n+1)/2+1}(0)}{(n+2)!}\prod_{j=0}^{(n+1)/2}(2j+1)-$$

$$-\frac{\mu''^{(n+1)/2+1}(0)}{(n+1)!}\prod_{j=0}^{(n-1)/2}(2j+1)=$$

$$=\left(2\frac{n+1}{2}+1\right)\frac{\mu''^{(n+1)/2+1}(0)}{(n+2)!}\prod_{j=0}^{(n-1)/2}(2j+1)-$$

$$-\frac{\mu''^{(n+1)/2+1}(0)}{(n+1)!}\prod_{j=0}^{(n-1)/2}(2j+1)=0.$$

Таким образом, методом математической индукции показано, что все коэффициенты q_n разложения в ряд Маклорена функции $h(x)=\mu'(x)/\mu(x)$ с индексами $n\geq 2$ будут нулевыми.

3. Получаем, что исходным условиям леммы удовлетворяют решения дифференциального уравнения $\mu'(x)/\mu(x)=\alpha x$, где $\alpha=\mu''(0)$. Найдем их:

$$\frac{d\mu(x)}{\mu(x)}=\alpha x\,dx;$$

$$\ln(\mu(x))=\frac{\alpha x^2}{2}+c,$$

где $c=const(x)$; $\mu(x)=e^{(\alpha x^2)/2+c}$.

Поскольку $\mu(0)=1$, то константа $c=0$ и $\mu(x)=e^{\alpha x^2/2}$, т. е. является нормированной гауссианой и описывает семейство показательных функций от квадрата аргумента. Из условия существования максимума при $x=0$ получаем, что $\alpha=\mu''(0)<0$. Поскольку степенной ряд $\sum\limits_{n=0}^{\infty}b_nx^n$ имеет бесконечный радиус сходимости, то в силу своего аналитического характера функция $\mu(x)=e^{\alpha x^2/2}$ на всем своем носителе S.

Лемма доказана.

Литература

1. *Gonella L.* Proposal for a revision of the measure theory and terminology // Alta Frequenza, 1975. Vol. XLIV. No. 10.
2. *Destouches J. L., Fevrier P.* New trends in expressing results of measurements // IMEKO Colloquium Proceedings. — Budapest, 1980.
3. *Mari L.* Notes on Fuzzy set theory as a tool for the measurement theory // Fundamental metrology, measurement theory and education: XII IMEKO World Congress Proceedings. — Beijing, China. 1991. Vol. III. P. 70–74.
4. *Reznik L. K., Jonson W. C., Solopchenko G. N.* Fuzzy interval as a basis for measurement theory // NASA Conference NAFIPS'94 Proceedings. — San-Antonio, Texas, 1994. P. 405–406.
5. *Резник Л. К.* Математическое обеспечение обработки нечеткой информации экспериментатора в ИВК // Архитектура, модели и программное обеспечение ИИС и ИВК: Труды ВНИИЭП. — Л.: ВНИИЭП, 1983. С. 45–55.
6. *Reznik L.* Measurement result uncertainty evaluation: New soft approaches? // Мягкие вычисления и измерения: Сб. трудов междунар. научн. конф. SCM-1999. — СПб., 1999. С. 21–24.
7. *Брусакова И. А.* Neuro и Fuzzy информационно-измерительные технологии для анализа априорных знаний интеллектуальных измерительных средств // Мягкие вычисления и измерения: Сб. трудов междунар. научн. конф. SCM-2003. — СПб., 2003. С. 27–32.

8. Руководство по выражению неопределенности измерения / Пер. с англ. под ред. В. А. Слаева. — СПб.: ВНИИМ им. Д. И. Менделеева, 1999.

9. МИ 1317-2004. Государственная система обеспечения единства измерений. Результаты и характеристики погрешности измерений. Формы представления. Способы использования при испытаниях образцов продукции и контроле их параметров: Издание официальное. — М.: Изд-во стандартов, 2004.

10. *Hung T. Nguen, Kreinovich V.* Nested intervals and sets: Concepts, relations to fuzzy sets, and applications // Application of interval computation / Eds. R. B. Kearfott, V. Kreinovich. — Dordrecht–Boston–London: Kluver Academic Publs., 1996. P. 245–290.

11. *Борисов А. Н., Алексеев А. В., Меркурьева Г. В. и др.* Обработка нечеткой информации в системах принятия решений. — М.: Радио и связь, 1989. 304 с.

12. *Пономарев А. С.* Нечеткие множества в задачах автоматизированного управления и принятия решений. — Харьков: НТУ ХПИ, 2005. 232 с.

13. *Яхъяева Г. Э.* Нечеткие множества и нейронные сети. — М.: Интернет-университет информационных технологий, Бином. Лаборатория знаний, 2006. 316 с.

14. *Солопченко Г. Н.* Представление измеряемых величин и погрешностей измерений как нечетких переменных // Измерительная техника, 2007. № 2. С. 3.

15. *Hung T. N., Kreinovich V., Chin-Wang T., Solopchenko G. N.* Why two sigma? A theoretical justification // Soft computing in measurement and information acquisition / Eds. L. Reznik, V. Kreinovich. — Berlin–Heidelberg: Springer-Verlag, 2003. P. 10–22.

16. *Семенов К. К., Солопченко Г. Н.* Теоретические предпосылки реализации метрологического автосопровождения программ обработки результатов измерений // Измерительная техника, 2010. № 6. С. 9–14.

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2012. Т. 6. Вып. 2. С. 113–121

ОСОБЕННОСТИ СЕМАНТИЧЕСКОГО ПОИСКА ИНФОРМАЦИОННЫХ ОБЪЕКТОВ НА ОСНОВЕ ТЕХНОЛОГИИ БАЗ ЗНАНИЙ

М. М. Шарнин[1], И. П. Кузнецов[2]

Аннотация: Рассматривается система семантического поиска информации в больших массивах документов на естественном языке (ЕЯ). Поиск основан на использовании лингвистического процессора, обеспечивающего автоматическое выделение из текстов информационных объектов (именованных сущностей), их признаков, связей и участие в действиях. В результате формируются структуры знаний. Аналогичным образом формируется структура запроса. Поиск, называемый семантическим, обеспечивается за счет сопоставления таких структур, где учитываются связи объектов, а также их участие в событиях, действиях.

Ключевые слова: семантический поиск; семантико-ориентированный лингвистический процессор; извлечение знаний из текстов; база знаний

1 Введение

Одной из актуальных задач в области информационных технологий является поиск информации в больших массивах документов — текстов на *естественном языке*. Для многих профессиональных пользователей поиск определяется их задачами. Например, задачи следователей-аналитиков (из области «Криминалистика») непосредственно связаны с поиском фигурантов, их адресов, деяний, связей между фигурантами, поиском по приметам, поиском похожих фигурантов и происшествий и многим другим. Для поиска используются документы криминальной полиции, имеющие вид текстов на ЕЯ: сводки происшествий и др.

Другой пример — задачи кадровых агентств, где документами являются резюме людей, желающих получить работу. Такие резюме часто пишутся в свободной форме — в виде текстов на ЕЯ. В резюме даются анкетные данные, места учебы и работы с указанием периодов и организаций или учебных заведений и т. д. Задачи кадровых агентств — поиск лиц по запросам клиентов, которые часто задаются на ЕЯ.

Следует отметить, что профессиональных пользователей интересует определенного сорта информация, которая зависит от предметной области. В приведенных выше примерах это лица, где они работают (организации), кем (профессии), чем занимаются (служебные обязанности) или в каких событиях участвовали (деяния лиц) и т. д. Подобную информацию будем называть *информационными объектами* или просто *объектами* (другое название — *именованные сущности*).

Поиск информационных объектов — это самостоятельная задача. Типовые поисковые машины (Google, Яндекс и др.) ищут ресурсы, содержащие слова запроса. Они не учитывают семантическую составляющую — наличие объектов, их связи.

Для поиска объектов требуется предварительная формализация текстов на ЕЯ — выделение не только объектов, но и всего, что с ними связано. Возникают структуры знаний.

В данной статье рассматривается поиск, основанный на сопоставлении таких структур. *Поиск* осуществляется не на уровне слов, а на уровне структур знаний, и поэтому является *семантическим*.

В настоящее время проблема семантического поиска приобретает все большую актуальность. Следует отметить семантическую поисковую систему AskNet (http://asknet.ru), которая «автоматически выбирает смысловые ответы на запросы пользователя», систему Hakia (http://hakia.com), основанную на хранилище семантической информации и технологии ранжирования найденных текстов по смыслу, а также системы Wolfram Alpha, Powerset и др. Во многих из них рассматриваются смысловые связи между терминами, для представления которых разрабатываются специальные формализмы.

Цель данной статьи — описание технологии поиска информационных объектов на основе струк-

[1]Институт проблем информатики Российской академии наук, keywen1@mail.ru
[2]Институт проблем информатики Российской академии наук, igor-kuz@mtu-net.ru

тур знаний, для извлечения которых используется семантико-ориентированный лингвистический процессор [1–3]. Такой поиск не является универсальным (как в системах Яндекс, Google). Его организация требует настройки лингвистического процессора на выделение объектов в определенной предметной области. Набор таких объектов ограничен. Соответственно, система настраивается давать точные ответы на определенный круг запросов.

В основе семантических поисков лежит технология баз знаний (БЗ), разработанная в рамках проектов ИПИ РАН [4]. Она включает в себя формализацию текстов (извлечение структур знаний), формализмы представления и хранения знаний (выделенных «смысловых элементов») в БЗ, а также методы сопоставления запроса и имеющейся информации на уровне структур знаний.

Для организации соответствующего технологического комплекса требуются формализмы, которые должны обладать определенными свойствами: быть как можно более простыми (в синтаксическом плане), обладать высокими изобразительными возможностями для представления знаний и обеспечивать в широких пределах логико-лингвистическую обработку [1]. Данными свойствами обладает язык расширенных семантических сетей (РСС) и продукционный язык их обработки — ДЕКЛ. На этой основе разработан инструментальный комплекс, ориентированный на обработку структур знаний [5, 6]. Комплекс использован для построения класса семантико-ориентированных лингвистических процессоров, преобразующих тексты на ЕЯ в формализм РСС, организации на этой основе БЗ и для разработки множества прикладных программ, обеспечивающих идентификацию объектов, выявление имплицитной информации, преобразование представлений, семантический поиск, принятие экспертных решений и др. Все эти задачи дополняют одна другую и решаются на одном уровне — структур знаний [4].

Отметим, что структуры знаний в виде РСС автоматически отображаются на языке XML [7] и могут быть использованы для построения прикладных программ на различных языках программирования. Подобный подход при соответствующей технологической доработке может быть основой крайне перспективного направления информатики — «Семантического Интернета».

2 Особенности обработки в базе знаний

Технология семантического поиска объектов на уровне структур знаний была разработана при построении систем «Аналитик» и «Криминал». Последняя была создана для ГУВД г. Москвы [2, 4, 6]. Эти системы ориентированы на работу с текстами на ЕЯ в определенной предметной области. В частности, система «Криминал» ориентирована на работу с большими потоками документов в области криминальной полиции: сводками происшествий, справками по уголовным делам, обвинительными заключениями, записными книжками фигурантов и др. Тексты автоматически формализуется с помощью семантико-ориентированного лингвистического процессора. При этом выделяются информационные объекты (фигуранты, их приметы, адреса, телефоны, даты, оружие, автотранспорт со всеми атрибутами и др.), а также связи между ними и разного рода деяниями, событиями. Участие объектов в одном действии считается одним из видов связи. Более того, сами действия — это тоже информационные объекты, которые связываются с временем, местом, а также причинно-следственными и другими отношениями. В результате возникают сложные структуры. На основе каждого документа формируется семантическая сеть (РСС), называемая *содержательным портретом документа* [8, 9]. Такие портреты образует *базу предметных знаний*, которая запоминается, а сами портреты связываются с соответствующими текстами.

Семантические поиски идут на уровне структур БЗ и включают в себя логический анализ признаков, связей. Например, поиск ответа на запрос в свободной форме (т. е. на ЕЯ) обеспечивается путем сопоставления содержательного портрета, построенного на основе запроса, и содержимого БЗ, т. е. сводится к поиску соответствующей структуры в БЗ. При этом широко используются онтологии, представленные в виде РСС, а также дополнительная информация, которая характеризует поисковый объект или ситуацию, но которая дается в тексте в неявной форме — как имплицитная информация, которую нужно восстанавливать [10].

В данной статье в качестве примера использования технологии БЗ рассматриваются задачи поиска похожих происшествий и лиц (фигурантов). При поиске похожих происшествий учитываются все действия и объекты, составляющие данное происшествие. При поиске похожего фигуранта учитывается только то, что связано с фигурантом. Эти задачи относятся к наиболее важным в области криминальной полиции. Они необходимы для идентификации лиц, установления их связей, порождения и проверки различных гипотез, планирования следственных действий. В данной статье рассматриваются методики и алгоритмы решения этих задач на структурном уровне, т. е. на основе различных видов связей с учетом особенностей

описываемых объектов, событий, происшествий. Ориентация сделана на использование семантических связей, а также методов логического анализа и нечеткого вывода. Отметим, что подобные методики использованы для семантического поиска других информационных объектов.

Задача поиска похожих происшествий и фигурантов решалась в рамках логико-аналитической системы «Криминал» с учетом ее задач и особенностей [4, 6].

В системе «Криминал» онтологии представлены в виде РСС и образуют *онтологическую базу*, которая находится в отдельном файле и объединяется с БЗ в процессе поиска. Онтологическая база определяет семантическое пространство терминов и признаков — с учетом их смысловой близости, синонимии и взаимоотрицания. За счет этого расширяется пространство поиска, повышается точность и надежность результатов, обеспечивается достаточная свобода использования слов и терминов в запросах и заданиях системе.

Все документы и полученные на их основе структуры знаний (содержательные портреты) помещаются в собственную базу данных, ориентированную на большие потоки информации и обеспечивающую их быстрый выбор — за счет индексных файлов (базы данных служат для хранения документов и структур знаний). Эти структуры по мере необходимости подкачиваются в оперативную память и вместе с онтологической базой образуют *оперативную базу знаний* (ОБЗ), где и осуществляется поиск. При этом допускается наличие множества баз данных (со своими БЗ) на различных компьютерах, связанных в сеть. Таким образом обеспечивается работа распределенных БЗ.

3 Содержательные портреты документов

Сеть (РСС), представляющая объекты и связи документа, образует его содержательный портрет, где все слова представлены в канонической форме. Такие портреты служат основой для семантического поиска.

Пример 1. Типовой документ (с номером 221) из сводок происшествий: *1.05.98 г. в 7.10 Фирсова Владимира Николаевича 1953 г.р. прож.: ул. Глаголева 25-1-273, работает АОЗТ «ХДУ», зам. директора, о том, что 1-05-98 г. неизвестные от д. 22 кор. 3 по ул. Тухачевского, похитили а/м ГАЗ 31029, черная, 1995 г/в, дв. 402-0019476. . .*

Его содержательный портрет имеет вид:

ДОК_(221,‘TEXT_98.TXT’,‘S_CRI.NL’)
ДАТА_(#1.5.1998,1998,МАЙ,∼1,7.1/4+)
FIO(ФИРСОВ,ВЛАДИМИР,НИКОЛАЕВИЧ,1953/5+)
АДР_(УЛ.,ГЛАГОЛЕВА,25,1,273/6+)
ПРОЖ.(5-,6-/7+)
ОРГ_(АОЗТ,ХДУ/8+)
РАБ_(5-,8-,ЗАМ.,ДИРЕКТОР/9+)
FIO(“ ”,“ ”,“ ”,НЕСКОЛЬКО/10+) НЕИЗВЕСТНЫЙ(10-)
АВТО_(ГАЗ,31029,ЧЕРНЫЙ,1995,Г/В,ДВ.,402-0019476/11+)
УГНАТЬ(10-,11-/12+)
ДАТА_(#1.5.1998,1998,МАЙ,∼1/14+)
КОГДА(12-,14-)
АДР_(УЛ.,ТУХАЧЕВСКОГО,ДОМ,22,КОРП.,3/15+)
ГДЕ(12-,15-)
ПРЕДЛ_(221,4-,5-,6-,8-,9-,О,ТОМ,12-,14-,15-)

Первый фрагмент ДОК_(221,‘TEXT_98.TXT’, ‘S_CRI.NL’) указывает на то, что содержательный портрет построен на основе документа 221 из файла ‘TEXT_98.TXT’. При этом были использованы лингвистические знания ‘S_CRI.NL’. Второй фрагмент представляет дату. Третий фрагмент представляет фигуранта — *Фирсова Владимира Николаевича*. Ему сопоставлен внутренний код 5+, с помощью которого представлено, где он проживает — ПРОЖ.(5-,6-/7+), где «5-» — код адреса. Здесь же представлены и другие объекты — организация (ОРГ_), место работы (РАБ_), автотранспорт (АВТО_) и др. Фрагмент УГНАТЬ(10-,11-/12+) представляет действие неизвестного лица (с кодом 10+), который *похитил* (= *угнал*) автомашину (с кодом 11+). Последний фрагмент ПРЕДЛ_(221,. . .) содержит коды других фрагментов и представляет порядок расположения соответствующей информации в тексте документа.

Такие портреты (в виде РСС) запоминаются в БЗ. Поиск сводится к сопоставлению таких портретов — запроса и содержимого БЗ [4, 6]. При поиске похожих фигурантов и происшествий важную роль играют не только объекты, но и действия типа УГНАТЬ и др. Помимо этого используется дополнительная информация, представленная в виде аналитических фрагментов (см. разд. 4).

4 Оценка документа по ключевым позициям

При организации семантических поисков важную роль играют признаки, задающие общий характер происшествий (*способы проникновения*, *совершения преступления* и др.), особенности фигуранта (их приметы) или особенности любого другого информационного объекта. Они могут в явном виде не присутствовать в тексте и требуют специального логического анализа для их выявления. С этой целью в процессе ввода документов с их формализацией производится оценка документа по ключевым

позициям. Она необходима для быстрого и качественного поиска, а также для выдачи информации в сжатом виде и объяснения результатов.

Оценка документа по ключевым позициям осуществляется на уровне структур знаний с помощью специальной программы постлингвистической обработки, реализующей идеологию семантических фильтров [3, 8, 9]. Оценка заключается в выделении особенностей описанного в документе происшествия (или особенностей какого-либо информационного объекта) и его соотнесении с соответствующими ветвями типовых классификаторов, находящихся в онтологической базе. Такое соотнесение осуществляется автоматически — на основе анализа содержательного портрета документа.

В результате строятся так называемые *аналитические фрагменты*, которые представляют в сжатом виде наиболее значимую информацию об объекте или происшествии и которые дополняют содержательный портрет документа. Они играют важную роль при поиске и аналитической обработке.

Рассмотрим примеры работы программы постлингвистической обработки на документах из области криминальной полиции.

Пример 2. Формирование по тексту описания словесного портрета фигуранта.

Основные классы онтологической базы, характеризующие фигурантов (лиц): *пол*, *возраст*, *рост*, *особые приметы*, *индивидуальные особенности*, *телосложение*, *тип лица*, *волосы*, *глаза*, *лоб*, *брови*, *нос*, *рот*, *губы*, *зубы*, *подбородок*, *уши*, *одежда*.

Текст на входе:

. . . На вид 45 лет, рост 170–175 см, полного т/сл., одет в рыжую лохматую шапку, зеленый пуховик, черные брюки, зимние ботинки коричневого цвета. . .

На выходе — аналитический фрагмент, представляющий в формализованном виде следующую информацию:

ВОЗРАСТ: 45,
РОСТ: 170, 175,
ТЕЛОСЛОЖЕНИЕ: ПОЛНЫЙ,
ОДЕЖДА: ШАПКА (РЫЖИЙ, ЛОХМАТЫЙ),
ПУХОВИК (ЗЕЛЕНЫЙ), БРЮКИ (ЧЕРНЫЙ),
БОТИНОК (ЗИМНИЙ, КОРИЧНЕВЫЙ).

Каждое слово с двоеточием представляет класс. Далее следуют подклассы. Слова в скобках поясняют или уточняют эти подклассы.

Подобное формализованное описание играет роль реферата. Оно строится по аналитическому фрагменту автоматически с помощью обратного лингвистического процессора. В данном случае программа постлингвистической обработки осуществляет автоматическое построение словесного портрета по тексту описания с его формализацией.

Пример 3. Выявление из текста описания основных характеристик происшествия.

Основные классы онтологической базы, характеризующие криминальные происшествия: *предварительные действия*, *способ проникновения*, *способ совершения преступления*, *преступные действия*, *предлог*, *организация*, *оружие*, *транспортные средства*, *ценные бумаги*, *драгоценные изделия*, *ценные изделия*.

Текст на входе:

. . . Найдена а/м ВАЗ 2109 темно-вишневого цвета г. н. К 939 ЕМ 70, в которой на передних сиденьях находятся два трупа мужчин кавказской национальности на вид 30–35 лет. Исследование показало, что смерть данных лиц наступила от огнестрельного ранения. На месте преступления были найдены и изъяты стреляные 2 гильзы от пистолета ТТ и 5 стреляных гильз, пуля от ПМ.

На выходе — аналитический фрагмент, представляющий в формализованном виде следующую информацию:

Преступные действия: РАНЕНИЕ (ОГНЕСТРЕЛЬНЫЙ),
ЛИЧНОСТЬ: НАЦИОНАЛЬНОСТЬ (Лицо кавказской национальности),
ОРУЖИЕ: ПИСТОЛЕТ (ТТ,ПМ),
АВТОМАШИНА: ВАЗ.

Подобное описание (как и в предыдущем примере) строится автоматически и играет роль сжатого описания или реферата.

5 Этапы поиска

Поиск похожих происшествий и фигурантов (как и других информационных объектов) осуществляется по запросам и заключается в анализе содержательных портретов документов на предмет их совпадения с содержательным портретом запроса [4, 6, 7]. Анализ осуществляется на уровне структур знаний, находящихся в оперативной БЗ. Вначале выделяются объекты запроса (например, фигуранта), их признаки — *приметы*, *деяния*, а также связанные с ними *адреса*, *телефоны*, *машины* и др. Анализ сводится к проверке наличия в документах объектов (фигурантов) с аналогичными признаками и связями. При этом используются следующие признаки и связи:

- первичные признаки (значимые слова запроса в каноническом виде);
- вторичные признаки (близкие по смыслу слова, уточняющие слова и др.), порожденные первичными признаками за счет информации онтологической базы;

– аналитические признаки (*способ проникновения*, *способ совершения преступления* и др.), взятые из аналитических фрагментов;
– свойства объектов (например, для фигурантов — *неизвестный*, *потерпевший*, *безработный*, для действий и событий — *время*, *место*);
– участие объектов в действиях.

Отметим, что в качестве запроса может быть взят любой документ или словесный портрет фигуранта. Тогда вначале будет сформирован содержательный портрет, в котором будут все объекты запроса. Далее пользователь может выбрать любой из них. Тогда система на содержательном уровне будет искать похожие объекты. Если пользователь выберет документ, то будет инициирован поиск похожих происшествий. Поиск является нечетким, так как не требуется полного совпадения слов-признаков. Находится только то, что является общим и объединяет запрашиваемый и найденный объекты. Это важно, так как точный поиск часто не дает результата.

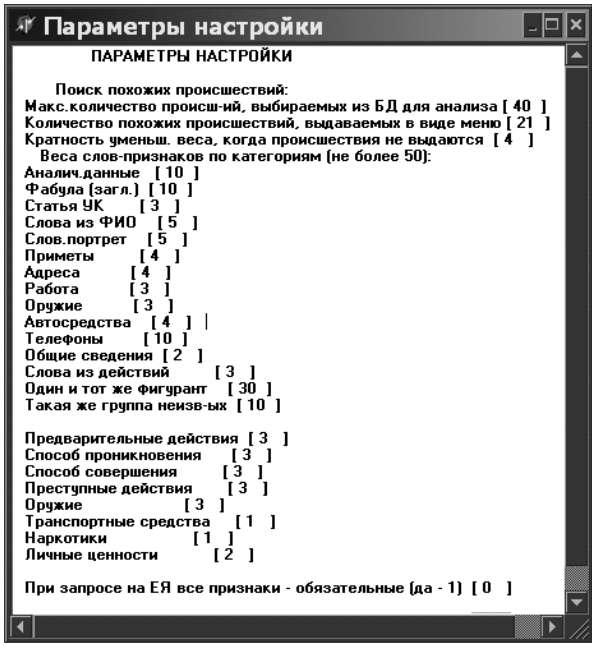
Этапы поиска похожих происшествий и фигурантов

Первый этап. Выделение значимых слов-признаков из содержательного портрета запроса с присвоением им весов. Они образуют первичные признаки. Под значимыми понимаются слова, которые не являются предлогами, союзами, понятиями широкого объема, вспомогательными глаголами и др. Напомним, что значимые слова блоком морфологического анализа приводятся в каноническую форму [2].

Если стоит задача поиска похожих происшествий, то используются все значимые слова запроса, дополненные аналитическими признаками. Если решается задача поиска похожих фигурантов, то из запроса выделяются только те части, которые относятся к указанному фигуранту: его ФИО, а также *приметы*, *адрес*, *деяния* и т. д. Только из этих частей берутся слова, которые в дальнейшем будут играть роль первичных признаков.

Выделенным словам-признакам присваиваются веса в зависимости от уникальности слова и вида информации (куда отнесено слово — к приметам, адресам и др.). Наибольшие веса присваиваются аналитическим признакам, относящимся к характеру происшествия и к фигурантам, описанным в запросе.

Отметим, что веса фактически отражают степень значимости той или иной информации при анализе степени сходства. В системе «Криминал» имеются специальные настроечные фреймы, дающие возможность пользователю изменять веса аналитических данных (*способ проникновения*, *оружие* и т. п.) и категорий (*приметы*, *адреса* и т. п.). Соответственно, будут меняться веса аналитических и других признаков. Таким способом акцентируется внимание на определенных моментах. Пример настроечного фрейма, задающего веса при поиске похожих происшествий (веса получены экспериментальным путем):



Аналогичный вид (с другими категориями и весами) имеет настроечный фрейм, определяющий веса слов-признаков при поиске похожих фигурантов.

Например, если дать высокий вес классу «*способ проникновения*» то система будет присваивать высокие веса происшествиям (документам) со способом проникновения, описанным в запросе. Если дать высокий вес адресам, то большой вес будет присваиваться документам с такими же улицами, номерами домов и квартир, как в запросе.

Второй этап. Дополнение набора признаков за счет онтологической базы. Это необходимо для расширения пространства поиска. Нужно учесть различные способы и средства описания того, что есть в запросе. На базе имеющихся признаков запроса порождаются вторичные признаки:

– близкие по смыслу термины (на основе фрагментов NEAR);
– поясняющие термины (на основе фрагментов SUB);
– противоречивые признаки (на основе фрагментов OR_OR).

Примеры фрагментов, взятых из онтологической базы:

NEAR(АТЛЕТИЧЕСКИЙ,МОГУЧИЙ,МОЩНЫЙ,
БОГАТЫРСКИЙ,СПОРТИВНЫЙ,К РЕПКИЙ)
SUB(ДОКУМЕНТ,ПАСПОРТ)
SUB(ДОКУМЕНТ,УДОСТОВЕРЕНИЕ)
SUB(ДОКУМЕНТ,“Водительские права”)
SUB(ДОКУМЕНТ,“Воинский билет”)
SUB(ДОКУМЕНТ,МЕТРИКА)
SUB(ДОКУМЕНТ,ПРОПУСК)
OR_OR(МОЛОДОЙ,“средний возраст”, ПОЖИЛОЙ)
NEAR(ПОЖИЛОЙ,СТАРЫЙ)
. . .

Поясним роль этих фрагментов на примерах. Если в запросе встретился признак БОГАТЫРСКИЙ (относящийся к фигуранту), то за счет фрагмента NEAR будут сформированы вторичные признаки: АТЛЕТИЧЕСКИЙ, МОГУЧИЙ, МОЩНЫЙ, СПОРТИВНЫЙ, КРЕПКИЙ, которые будут принимать участие при поиске.

Если в запросе встретился термин ДОКУМЕНТ, то за счет фрагментов типа SUB будут сформированы способы его расшифровки: это может быть ПАСПОРТ, УДОСТОВЕРЕНИЕ, «Водительские права», «Воинский билет», МЕТРИКА, ПРОПУСК. Они будут также учитываться при поиске.

Если в запросе фигурант был охарактеризован как ПОЖИЛОЙ (это могла сделать и сама система путем анализа возраста), то за счет фрагментов типа OR_OR будут сформированы опровергающие признаки: МОЛОДОЙ, «средний возраст», которые используются при оценке степени сходства со знаком минус.

Вторичным признакам также присваиваются веса — в зависимости от веса признака, который их породил.

При наличии в запросе фигурантов производится анализ их ФИО. Отметим, что полные имена и отчества при построении содержательного портрета уже преобразуются к единому виду — в каноническую форму. На данном этапе на их основе порождаются инициалы, которые тоже играют роль признаков. И наоборот, по инициалам порождаются возможные имена и отчества. Это позволяет при поиске более полно охватить возможные случаи написания ФИО.

Третий этап. Быстрый поиск (по индексным файлам) в базах данных содержательных портретов документов с указанными признаками. Поиск осуществляется на основе выделенных из запроса слов-признаков и заключается в подсчете взвешенной суммы весов совпавших признаков. В качестве результата выдаются номера найденных документов — в порядке взвешенных сумм.

При наличии в запросе фигурантов с ФИО производится дополнительный поиск документов, при котором обязательными признаками делаются пары: полные фамилия и имя или полные имя и отчество. Словом, находятся документы, где есть и то, и другое. Это позволяет избежать потерь при поиске и идентификации фигурантов.

Четвертый этап. Подкачка из базы данных в оперативную память семантических сетей — содержательных портретов документов с наибольшими взвешенными суммами. В результате (вместе с онтологической базой) образуется ОБЗ, представляющая собой большую семантическую сеть, доступную для быстрого выполнения сложных операций сравнения и логического анализа.

Пятый этап. Детальный анализ на совпадение слов-признаков, связанных с объектами (или выбранным объектом) запроса и объектами, находящимися в ОБЗ. При этом сравнение идет по категориям: ФИО фигуранта из запроса сравнивается с ФИО фигурантов из ОБЗ, связанные с ними приметы сравниваются с приметами, свойства — со свойствами, действия — с действиями и т. д. В результате находятся похожие объекты. Подсчитывается их вес, который определяет степень сходства с объектами (или объектом) запроса. Выбираются объекты с наибольшими весами. При этом учитываются следующие факторы:

- веса совпавших порожденных аналитических признаков, определяющих характер объекта или особенности фигуранта;
- веса совпавших слов-признаков (в том числе вторичных);
- соотнесенность признаков к той или иной категории;
- связь признаков, заданная в онтологической базе (близкие по смыслу или поясняющие);
- сильное совпадение по какой-либо категории признаков (например, совпадает большинство примет);
- наличие противоречивых признаков.

Каждое совпадение дополняет общий вес выбранного объекта — к нему добавляется вес совпавшего признака. При наличии противоречивых признаков их веса вычитаются.

При анализе чисел и интервалов на их совпадение (например, ВОЗРАСТ, РОСТ, номер дома, квартиры, год и др.) рассматриваются различные варианты:

- равенство чисел;
- число входит в интервал;
- пересечение интервалов;
- близость числовых значений.

В зависимости от варианта совпадения и от категории (приметы, адрес, время и др.) к общему весу документа добавляется определенная величина.

При поиске похожих происшествий в первую очередь учитывается сходство аналитических признаков и криминальных действий, а уже затем объектов, участвующих в этих действиях. Для этого категориям присваиваются соответствующие веса. Общий вес выбранного (из ОБЗ) происшествия подсчитывается как сумма весов входящих в него признаков и объектов (действие — это тоже объект).

При поиске фигурантов различается два случая.

Первый случай — когда в запросе заданы ФИО фигуранта. Тогда в ОБЗ производится поиск фигурантов с аналогичными ФИО. При этом учитываются случаи совпадения инициалов с полными именами или отчествами (такое совпадение дает меньший вес). Подсчитывается общая степень совпадения — в зависимости от совпавших признаков.

Множество найденных фигурантов с высокими весами является основой для дальнейшего анализа. К ним добавляются веса, полученные от совпадения свойств, а также от совпадения связанных с фигурантами примет, адресов, телефонов и др. В результате находятся фигуранты с высокими весами, отражающими степень сходства с лицом, описанным в запросе.

Если в ОБЗ не найдено фигурантов с ФИО, заданными в запросе, то ищутся фигуранты, у которых может быть другое имя или отчество. Возникают противоречивые признаки, которые уменьшают вес анализируемого фигуранта. При этом акцент смещается на сравнение связей.

Второй случай — когда запрашивается неизвестное лицо (фигурант). Тогда поиск и сравнение идет по связанным с этим лицом приметам, действиям, адресам и другим объектам. В ОБЗ ищутся лица с аналогичными связями.

При поиске похожих происшествий найденные фигуранты с их степенями совпадения запоминаются, а сами степени дополняют вес соответствующего документа. Помимо этого учитываются веса других совпавших признаков. В результате находятся происшествия (документы) с высокими весами, отражающими степень сходства с запросом.

Шестой этап. Выдача похожих происшествий или фигурантов, упорядоченных по степени сходства, в виде списка или меню.

Седьмой этап. Выдача объяснений. Пользователь может выбрать из упомянутого меню любой пункт, соответствующий происшествию или фигуранту. Система на основе совпавших признаков формирует объяснение сходства в виде понятного текста на русском языке.

6 Выдача и объяснение результатов

Как отмечалось ранее, вся обработка в системе «Криминал» осуществляется на уровне семантических сетей в рамках специально созданного для этого инструментария языка — ДЕКЛ [5, 6]. Находятся фрагменты семантической сети, представляющие похожие происшествия или фигурантов с совпавшими признаками. При выдаче соответствующего меню и объяснении результатов такие фрагменты преобразуются на понятный пользователю язык — естественный. Это делается с помощью обратного лингвистического процессора.

При формировании меню формируются краткие описания происшествий или фигурантов. При объяснении результатов (когда пользователь выбирает из меню интересующее его происшествие или фигуранта) дается краткое описание выбранного происшествия (фигуранта), указываются совпавшие и противоречивые признаки, а также дается сам текст описания. Этого достаточно, чтобы помочь пользователю самому оценить степень сходства или адекватности запросу.

Пример 4. Проиллюстрируем сказанное на примере выдачи результатов поиска похожих криминальных происшествий и похожих фигурантов.

Текст на входе (взят из документа с номером 63): *. . . На лестничной площадке 3-го этажа двое неизвестных из неустановленного оружия нанесли два сквозных ранения в голову и живот Лихомову Владимиру Ивановичу, 1954 г.р., неработающий, прож.: Тюменская. . . С места происшествия изъято: 1 пуля и 1 гильза калибра 7.62 мм предположительно от пистолета ТТ. . .*

Меню похожих происшествий выглядит следующим образом:

На документ 63 содержательно похожи:

- Док-т 1231 (БЗ-1) УБИЙСТВО 29.3.1996 (вес 142);
- Док-т 4323 (БЗ-1) ОГРАБЛЕНИЕ 20.6.1996 (вес 111);
- Док-т 81 (БЗ-2) УБИЙСТВО 1.7.1995 (вес 92);
- . . .

При выборе пункта 1 данного меню на экран будет выдано объяснение причин сходства документов 63 и 1231:

Похожее происшествие — 1231 из БЗ-1 (вес 142).

В происшествии 1231 встретились те же признаки:

Преступные действия: РАНЕНИЕ, ГОЛОВА.
Оружие: ПИСТОЛЕТ, ТТ.
Фабула: УБИЙСТВО.
Работа: НЕРАБОТАЮЩИЙ.
Действие: ИЗЪЯТЬ ГИЛЬЗА.
Общие сведения: РЕЗУЛЬТАТ, ПРОВЕДЕНИЕ, СОТРУДНИК, ОКАЗАТЬСЯ, КАЛИБР.
<Текст документа 1231 из БЗ-1>.

Пример 5. Меню похожих фигурантов выглядит следующим образом:

На фигуранта ЛИХОМОВ ВЛАДИМИР ИВАНОВИЧ похожи:

– ЛИХОМОВ ВЛАДИМИР ПЕТРОВИЧ 1943, док. 4437 из БЗ-1 (вес 42);
– без ФИО в кол-ве 1, док. 81 из БЗ-1 (вес 35);
– КОВАЛЕВ ВЛАДИМИР ИВАНОВИЧ 1956, док. 24 из БЗ-2 (вес 30);
– . . .

При выборе пункта 1 данного меню на экран будет выдано объяснение причин сходства фигурантов ЛИХОМОВ ВЛАДИМИР ИВАНОВИЧ и ЛИХОМОВ ВЛАДИМИР ПЕТРОВИЧ:

Похожий фигурант (вес 44) — ЛИХОМОВ ВЛАДИМИР ПЕТРОВИЧ 1943.
Особые приметы: ОТМЕТИНА (РАНЕНИЕ),
ТЕЛОСЛОЖЕНИЕ: ТОЛСТЫЙ,
СТАТУС: ПОТЕРПЕВШИЙ (РАНЕНИЕ).

У фигуранта встретились те же признаки:

ТЕЛОСЛОЖЕНИЕ: ТОЛСТЫЙ,
СТАТУС: ПОТЕРПЕВШИЙ РАНЕНИЕ,
Работа: НЕРАБОТАЮЩИЙ.
ФИО: ЛИХОМОВ ВЛАДИМИР.
Не совпадают ФИО: 1943 (*было* 1954).
Не совпадают ФИО: ПЕТРОВИЧ (*было* ИВАНОВИЧ).
Входит в документ с номером 4437 из БЗ-2.
<Текст документа 4437>.

Отметим некоторые важные моменты.

Во-первых, в содержательных портретах представлены объекты и действия. Их сопоставление играет важную роль при поиске похожих происшествий, при классификации лиц как *потерпевших* или *преступников* и во многих других случаях.

Во-вторых, дается оценка документа по ключевым позициям, представляющая в сжатом виде наиболее значимую информацию и играющая роль реферата.

В-третьих, при анализе степени сходства запроса и документа используются признаки типа «*Преступные действия*», «*Угроза оружием*» и др., которые в явном виде могут отсутствовать в тексте и которые выявляются системой в процессе постлингвистической обработки. Соответственно, такие признаки вводятся в объяснения.

В-четвертых, допускается поиск похожих фигурантов без ФИО (поиск неизвестных лиц) по связанной информации, например по словесному портрету.

И последнее: использование привычных человеку классификаторов (они представлены в онтологической базе) делает результат реферирования и объяснения более понятным.

7 Заключение

Особенность предлагаемых в данной статье методик и алгоритмов семантического поиска состоит в следующем:

(1) вся обработка осуществляется на уровне структур знаний, т. е. содержательных портретов документов. Они образуют БЗ, которая вместе с правилами преобразования (продукциями языка ДЕКЛ) образует законченный технологический комплекс, ориентированный на сложные задачи, связанные с логическим выводом, преобразованием представлений, экспертными решениями. В результате обеспечивается анализ высокой степени глубины и сложности;

(2) выделяются и используются разнообразные признаки. Учитывается наличие множества объектов (лиц, телефонов, оружия и т. п. — до 40 типов) и аналитических признаков, характеризующих происшествия и фигурантов. Для расширения пространства поиска используется онтологическая база;

(3) допускается работа с многими БЗ, связанными через сеть или Интернет. Они образуют распределенную БЗ.

Описанные методики и алгоритмы семантического поиска реализованы в рамках систем «Аналитик», «Криминал», «Поток» и апробированы при работе с различными корпусами текстов, среди которых: сообщения СМИ, сводки происшествий, обвинительные заключения, записные книжки фигурантов и др. Эти методики использованы в различных приложениях и показали высокую степень универсальности. В перспективе они могут послужить основой для создания комплекса поисковых программ, составляющих «Семантический Интернет».

Литература

1. *Кузнецов И. П.* Семантические представления. — М.: Наука, 1986. 290 с.
2. *Кузнецов И. П., Кузнецов В. П., Мацкевич А. Г.* Система выявления из документов значимой информации на основе лингвистических знаний в форме семантических сетей // Диалог-2000: Труды Междунар. семинара по компьютерной лингвистике и ее приложениям. — Протвино, 2000. Т. 2.
3. *Kuznetsov I., Matskevich A.* System for extracting semantic information from natural language text // Диалог-2002: Труды Междунар. семинара по компьютерной лингвистике и ее приложениям (Протвино). — М.: Наука, 2002. Т. 2.
4. Лаборатория компьютерной лингвистики ИПИ РАН: Официальный сайт. www.IpiranLogos.com.
5. *Кузнецов И. П., Шарнин М. М.* Продукционный язык программирования ДЕКЛ // Система обработки декларативных структур знаний Деклар-2. — М.: ИПИ РАН, 1988.
6. *Кузнецов И. П., Мацкевич А. Г.* Семантико-ориентированные системы на основе баз знаний. — М.: МТУСИ, 2007. 173 с.
7. *Kuznetsov I. P., Kozerenko E. B.* Linguistic processor Semantix for knowledge extraction from natural texts in Russian and English // ICAI 2008: 2008 Conference (International) on Artificial Intelligence Proceedings. — Las Vegas: CSREA Press, 2008. P. 835–841.
8. *Кузнецов И. П., Мацкевич А. Г.* Особенности организации базы предметных и лингвистических знаний в системе АНАЛИТИК // Диалог-2003: Труды Междунар. конф. по компьютерной лингвистике и интеллектуальным технологиям. — Протвино, 2003. С. 373–378.
9. *Кузнецов И. П., Мацкевич А. Г.* Англоязычная версия системы автоматического выявления значимой информации из текстов естественного языка // Диалог-2005: Труды Междунар. конф. по компьютерной лингвистике и интеллектуальным технологиям (Звенигород). — М.: Наука, 2005. С. 303–311.
10. *Kuznetsov I. P.* Identifying role functions of persons on the basis of knowledge structures) // Компьютерная лингвистика и интеллектуальные технологии: По матлам ежегодной Междунар. конф. «Диалог 2011». — М.: РГГУ, 2011. Вып. 10(17). С. 391–402.

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2012. Т. 6. Вып. 2. С. 122–128

О СКОРОСТИ СХОДИМОСТИ ОЦЕНКИ РИСКА ПОРОГОВОЙ ОБРАБОТКИ ВЕЙВЛЕТ-КОЭФФИЦИЕНТОВ К НОРМАЛЬНОМУ ЗАКОНУ ПРИ ИСПОЛЬЗОВАНИИ РОБАСТНЫХ ОЦЕНОК ДИСПЕРСИИ*

О. В. Шестаков[1]

Аннотация: Исследуются асимптотические свойства оценки риска при пороговой обработке коэффициентов вейвлет-разложения функции сигнала. Получены некоторые оценки скорости сходимости распределения оценки риска к нормальному закону.

Ключевые слова: вейвлеты; пороговая обработка; оценка риска; нормальное распределение; оценка скорости сходимости

1 Введение

Методы обработки сигналов и изображений с помощью вейвлет-разложения применяются в самых разнообразных областях, включая геофизику, физику плазмы, вычислительную томографию, компьютерную графику и т. д. Основные задачи, для решения которых используется вейвлет-разложение, — это сжатие сигналов/изображений и удаление шума. При этом строится оценка сигнала или изображения, основанная на пороговой обработке вейвлет-коэффициентов, которая обнуляет коэффициенты, не превышающие заданного порога. Наличие шума неизбежно приводит к погрешностям в оцениваемом сигнале/изображении. Свойства оценки таких погрешностей (риска) исследовались в работах [1–8]. При определенных условиях оценка риска является состоятельной и асимптотически нормальной [6]. В данной работе исследуется вопрос о скорости сходимости распределения оценки риска к нормальному закону в одномерном случае (т. е. при обработке одномерных сигналов).

Введем необходимые понятия и обозначения. При использовании вейвлет-разложения функция $f \in L^2(\mathbf{R})$, описывающая сигнал, представляется в виде ряда из сдвигов и растяжений некоторой вейвлет-функции ψ:

$$f = \sum_{j,k \in Z} \langle f, \psi_{j,k} \rangle \psi_{j,k} \,, \tag{1}$$

где $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$ (семейство $\{\psi_{j,k}\}_{j,k\in Z}$ образует ортонормированный базис в $L^2(\mathbf{R})$). Индекс j в (1) называется масштабом, а индекс k — сдвигом. Функция ψ должна удовлетворять определенным требованиям [9], однако ее можно выбрать таким образом, чтобы она обладала некоторыми полезными свойствами, например была дифференцируемой нужное число раз и имела заданное число M нулевых моментов [9], т. е.

$$\int_{-\infty}^{\infty} x^k \psi(x)\, dx = 0 \,, \quad k = 0, \ldots, M-1 \,.$$

В дальнейшем будут рассматриваться функции $f \in L^2(\mathbf{R})$ с носителем на отрезке $[0,1]$, равномерно регулярные по Липшицу с некоторым параметром $\gamma > 0$, т. е. такие функции, для которых существует константа $L > 0$ и полином P_y степени $n = \lfloor \gamma \rfloor$ такой, что для любого $y \in [0,1]$ и любого $x \in \mathbf{R}$

$$|f(x) - P_y(x)| \leqslant L\,|x-y|^\gamma \,.$$

Для этих функций f известно [10], что если вейвлет-функция M раз непрерывно дифференцируема ($M \geqslant \gamma$), имеет M нулевых моментов и быстро убывает на бесконечности вместе со своими производными, т. е. для всех $0 \leqslant k \leqslant M$ и любого $m \in N$ найдется константа C_m, для которой при всех $x \in \mathbf{R}$

$$\left|\psi^{(k)}(x)\right| \leqslant \frac{C_m}{1+|x|^m} \,,$$

то найдется такая константа $A > 0$, что

$$\langle f, \psi_{j,k} \rangle \leqslant \frac{A}{2^{j(\gamma+1/2)}} \,. \tag{2}$$

*Работа выполнена при финансовой поддержке РФФИ (гранты 11-01-00515 и 11-01-12026-офи-м).

[1]Московский государственный университет им. М. В. Ломоносова; Институт проблем информатики Российской академии наук, oshestakov@cs.msu.su

На практике функции сигнала всегда заданы в дискретных отсчетах. Без ограничения общности будем считать, что функция f задана в точках i/N ($i = 1, \ldots, N$, где $N = 2^J$ для некоторого J): $f_i = f(i/N)$. Дискретное вейвлет-преобразование представляет собой умножение вектора значений функции f (обозначим его через $\overline{f}$) на ортогональную матрицу W, определяемую вейвлет-функцией ψ [10]: $\overline{f}^W = W\overline{f}$. При этом если перейти к двойному индексу (j,k), то дискретные вейвлет-коэффициенты будут связаны с непрерывными следующим образом: $f^W_{j,k} \approx \sqrt{N}\langle f, \psi_{j,k}\rangle$ (см., например, [2] или [11]). В дальнейшем для удобства будем нумеровать дискретные вейвлет-коэффициенты так же, как отсчеты функции f: одним индексом i вместо двойного индекса (j,k).

В реальных наблюдениях всегда присутствует шум. В данной работе рассматривается аддитивная модель шума: $Y_i = f_i + z_i$, $i = 1, \ldots, N$, где z_i — независимые случайные величины, имеющие нормальное распределение с нулевым средним и дисперсией σ^2. Тогда в силу ортогональности матрицы W для дискретных вейвлет-коэффициентов принимается следующая модель:

$$Y_i^W = f_i^W + z_i^W\,,\ \ i = 1, \ldots, N\,,$$

где z_i^W также независимы и нормально распределены с нулевым средним и дисперсией σ^2, а f_i^W равны соответствующим непрерывным вейвлет-коэффициентам, умноженным на $\sqrt{N}$.

2 Пороговая обработка и оценка риска

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

$$R_N(f) = \sum_{i=1}^{N} \mathrm{E}\left(f_i^W - \rho_T(Y_i^W)\right)^2\,.$$

Здесь присутствуют неизвестные величины f_i^W, поэтому вычислить значение $R_N(f)$ нельзя. Однако его можно оценить. В каждом слагаемом если $\left|Y_i^W\right| > T$, то вклад этого слагаемого в риск составляет $\sigma^2 + T^2$, а если $\left|Y_i^W\right| \leqslant T$, то вклад составляет $(f_i^W)^2$. Поскольку $\mathrm{E}(Y_i^W)^2 = \sigma^2 + (f_i^W)^2$, величину $(f_i^W)^2$ можно оценить разностью $(Y_i^W)^2 - \sigma^2$.

Таким образом, в качестве оценки риска можно использовать следующую величину:

$$\widetilde{R}_N(f) = \sum_{i=1}^{N} F[(Y_i^W)^2]\,, \tag{3}$$

где

$$F[x] = (x - \sigma^2)\mathbb{1}_{|x|\leqslant T^2} + (\sigma^2 + T^2)\mathbb{1}_{|x|>T^2}\,.$$

Для определенной таким образом оценки риска справедливо следующее утверждение [10].

Теорема 1. $\mathrm{E}\widetilde{R}_N(f) = R_N(f)$, *т. е.* $\widetilde{R}_N(f)$ *является несмещенной оценкой для* $R_N(f)$.

В работах [1, 3] было предложено использовать порог $T = \sigma\sqrt{2\ln N}$. Было показано, что при таком пороге риск близок к минимальному [1]. Этот порог получил название «универсальный». В дальнейшем будет использоваться именно такой вид порога.

Зачастую дисперсия σ^2 неизвестна, и ее также необходимо оценивать, при этом выражения (4) принимают вид

$$\widehat{R}_N(f) = \sum_{i=1}^{N} \widehat{F}[(Y_i^W)^2], \tag{4}$$

где

$$\widehat{F}[x] = (x - \hat{\sigma}^2)\mathbb{1}_{|x|\leqslant \hat{T}^2} + (\hat{\sigma}^2 + \hat{T}^2)\mathbb{1}_{|x|>\hat{T}^2}\,;$$

$$\hat{T} = \hat{\sigma}\sqrt{2\ln N}\,.$$

Обычно дисперсия σ^2 (или среднеквадратичное отклонение σ) оценивается по выборке сигнала, однако ее можно оценить и по независимой выборке. Для этого следует произвести измерение пустого сигнала, тогда наблюдения будут представлять собой чистый шум, по которому и оценивается σ^2. В следующих разделах будут рассмотрены оба случая.

3 Оценка скорости сходимости распределения оценки риска к нормальному закону

В работах [7, 6, 8] исследуется асимптотическое поведение оценки риска $\widehat{R}_N(f)$ при использовании

различных оценок $\hat{\sigma}$. Показано, что при определенных условиях оценка риска является асимптотически нормальной, и в случае использования выборочной дисперсии получены оценки скорости сходимости к нормальному закону. В этом разделе будут получены оценки скорости сходимости распределения $\widetilde{R}_N(f)$ к нормальному закону при использовании в качестве $\hat{\sigma}$ соответствующим образом нормированного интерквартильного размаха $\hat{\sigma}_R$ и абсолютного медианного отклонения от медианы $\hat{\sigma}_M$ в предположении, что эти оценки строятся по независимой выборке $(Y'_1, \ldots, Y'_N)$ из нормального распределения. Преимущество использования таких оценок заключается в их робастности, т. е. нечувствительности к выбросам [12, 13], и в полной мере проявляется, когда дисперсия оценивается по выборке сигнала. Оценки $\hat{\sigma}_R$ и $\hat{\sigma}_M$ определяются следующим образом:

$$\left.\begin{aligned} \hat{\sigma}_R &= \frac{Y'_{N,3/4} - Y'_{N,1/4}}{2\xi_{3/4}}; \\ \hat{\sigma}_M &= \frac{\underset{1\leqslant i\leqslant N}{\operatorname{med}} |Y'_i - \underset{1\leqslant j\leqslant N}{\operatorname{med}} Y'_j|}{\xi_{3/4}}, \end{aligned}\right\} \qquad (5)$$

где $Y'_{N,1/4}$ и $Y'_{N,3/4}$ — выборочные квантили порядка 1/4 и 3/4, $\xi_{3/4}$ — теоретическая квантиль порядка 3/4 стандартного нормального распределения, а med обозначает выборочную медиану.

Далее для удобства будем обозначать Y_i^W через X_i, а f_i^W через a_i.

Теорема 2. *Пусть f задана на отрезке $[0,1]$ и является равномерно регулярной по Липшицу с параметром $\gamma = 1/2 + \alpha$ ($\alpha > 0$) и пусть оценка $\hat{\sigma}$, равная $\hat{\sigma}_R$ или $\hat{\sigma}_M$, не зависит от наблюдений X_i, тогда существует такая константа C_0, что*

$$\sup_{x\in\mathbf{R}} \left| \mathrm{P}\left(\frac{\widehat{R}_N(f) - R_N(f)}{\sigma^2\sqrt{2N}} < x \right) - \Phi_\Sigma(x) \right| \leqslant \leqslant \frac{C_0 (\ln N)^{1/2+1/(4(\alpha+1))}}{N^{1/4-1/(4(\alpha+1))}}, \qquad (6)$$

где $\Phi_\Sigma(x)$ — функция распределения нормального закона с нулевым средним и дисперсией $\Sigma = 1 + 2[4\xi_{3/4}\phi(\xi_{3/4})]^{-2}$ ($\phi(x)$ — плотность стандартного нормального распределения). Константа C_0 зависит от α, A, σ и от того, какая из оценок используется — $\hat{\sigma}_R$ или $\hat{\sigma}_M$.

Доказательство. Поступая, как в работе [8], запишем разность $\widehat{R}_N(f) - R_N(f)$ в виде:

$$\widehat{R}_N(f) - R_N(f) = S_N + V_N,$$

где

$$S_N = \sum_{i=1}^N \left(X_i^2 \mathbb{1}_{|X_i|\leqslant\hat{T}} - \mathrm{E}X_i^2 \mathbb{1}_{|X_i|\leqslant T} \right) + 2\sum_{i=1}^N \left(\hat{\sigma}^2 \mathbb{1}_{|X_i|>\hat{T}} - \mathrm{E}\sigma^2 \mathbb{1}_{|X_i|>T} \right) + \sum_{i=1}^N \left(\hat{T}^2 \mathbb{1}_{|X_i|>\hat{T}} - \mathrm{E}T^2 \mathbb{1}_{|X_i|>T} \right);$$

$$V_N = N\left(\sigma^2 - \hat{\sigma}^2\right).$$

Рассмотрим S_N. Разобьем это слагаемое на три суммы U_N, W_N и Z_N:

$$U_N = \sum_{i\in I_1} \left(X_i^2 - \mathrm{E}X_i^2 \right),$$

$$W_N = -\sum_{i\in I_1} \left(X_i^2 \mathbb{1}_{|X_i|>\hat{T}} - \mathrm{E}X_i^2 \mathbb{1}_{|X_i|>T} \right) + 2\sum_{i\in I_1} \left(\hat{\sigma}^2 \mathbb{1}_{|X_i|>\hat{T}} - \mathrm{E}\sigma^2 \mathbb{1}_{|X_i|>T} \right) + \sum_{i\in I_1} \left(\hat{T}^2 \mathbb{1}_{|X_i|>\hat{T}} - \mathrm{E}T^2 \mathbb{1}_{|X_i|>T} \right);$$

$$Z_N = \sum_{i\in I_2} \left(X_i^2 \mathbb{1}_{|X_i|\leqslant\hat{T}} - \mathrm{E}X_i^2 \mathbb{1}_{|X_i|\leqslant T} \right) + 2\sum_{i\in I_2} \left(\hat{\sigma}^2 \mathbb{1}_{|X_i|>\hat{T}} - \mathrm{E}\sigma^2 \mathbb{1}_{|X_i|>T} \right) + \sum_{i\in I_2} \left(\hat{T}^2 \mathbb{1}_{|X_i|>\hat{T}} - \mathrm{E}T^2 \mathbb{1}_{|X_i|>T} \right),$$

где I_1 — множество тех i, для которых в силу (2) выполнено $|a_i| \leqslant A/(\ln N)^{1/2}$, а I_2 — множество остальных i. Оценим сумму $W_N + Z_N$.

При произвольном $\varepsilon > 0$, используя неравенство Чебышева, получаем:

$$\mathrm{P}\left(|W_N + Z_N| > \varepsilon\right) \leqslant \frac{\mathrm{E}|W_N| + \mathrm{E}|Z_N|}{\varepsilon}.$$

Рассмотрим $\mathrm{E}|Z_N|$:

$$\mathrm{E}|Z_N| \leqslant \sum_{i\in I_2} \mathrm{E}\left| X_i^2 \mathbb{1}_{|X_i|\leqslant\hat{T}} - \mathrm{E}X_i^2 \mathbb{1}_{|X_i|\leqslant T} \right| + 2\sum_{i\in I_2} \mathrm{E}\left| \hat{\sigma}^2 \mathbb{1}_{|X_i|>\hat{T}} - \mathrm{E}\sigma^2 \mathbb{1}_{|X_i|>T} \right| + \sum_{i\in I_2} \mathrm{E}\left| \hat{T}^2 \mathbb{1}_{|X_i|>\hat{T}} - \mathrm{E}T^2 \mathbb{1}_{|X_i|>T} \right|.$$

Можно показать, что

$$\xi_{3/4}\,|\hat{\sigma}_M - \hat{\sigma}_A| \leqslant \left| \underset{1\leqslant i\leqslant N}{\operatorname{med}} Y'_i \right| \text{ п. в.}, \qquad (7)$$

где

$$\hat{\sigma}_A = \frac{\underset{1\leqslant i\leqslant N}{\operatorname{med}} |Y_i'|}{\xi_{3/4}}\,.$$

Это соотношение позволяет использовать для абсолютного медианного отклонения многие результаты, справедливые для выборочных квантилей. Далее, пользуясь (7) и результатами работ [12, 14–17], можно показать, что

$$\mathrm{E}\hat{\sigma}_R^2 = \sigma^2 + O\left(\frac{1}{N^{3/4}}\right);\quad \mathrm{D}\hat{\sigma}_R = O\left(\frac{1}{N}\right); \qquad (8)$$

$$\mathrm{E}\hat{\sigma}_M^2 = \sigma^2 + O\left(\frac{1}{N^{1/2}}\right);\quad \mathrm{D}\hat{\sigma}_M = O\left(\frac{1}{N}\right). \qquad (9)$$

Поскольку f регулярна по Липшицу с $\gamma = 1/2+\alpha$, число слагаемых в каждой из трех сумм в $\mathrm{E}\,|Z_N|$ не превосходит $B_1(N\ln N)^{1/(2(\alpha+1))}$, где B_1 — некоторая константа, зависящая от α. Учитывая соотношения (8) и (9), можно показать, что слагаемые в первой и третьей суммах не превосходят $B_1'\sigma^2\ln N$, а слагаемые во второй сумме не превосходят $B_2'\sigma^2$ с некоторыми константами B_1' и B_2'. Следовательно, $\mathrm{E}\,|Z_N|$ не превосходит $B_2 N^{1/(2(\alpha+1))}(\ln N)^{1+1/(2(\alpha+1))}$ для некоторой константы B_2.

Оценим теперь $\mathrm{E}\,|W_N|$:

$$\begin{aligned}\mathrm{E}\,|W_N| &\leqslant \sum_{i\in I_1}\mathrm{E}\left|X_i^2\mathbb{1}_{|X_i|>\hat{T}} - \mathrm{E}X_i^2\mathbb{1}_{|X_i|>T}\right| + \\ &+ 2\sum_{i\in I_1}\mathrm{E}\left|\hat{\sigma}^2\mathbb{1}_{|X_i|>\hat{T}} - \mathrm{E}\sigma^2\mathbb{1}_{|X_i|>T}\right| + \\ &+ \sum_{i\in I_1}\mathrm{E}\left|\hat{T}^2\mathbb{1}_{|X_i|>\hat{T}} - \mathrm{E}T^2\mathbb{1}_{|X_i|>T}\right|. \qquad (10)\end{aligned}$$

Оценим первую сумму. Имеем:

$$\begin{aligned}&\mathrm{E}\left|X_i^2\mathbb{1}_{|X_i|>\hat{T}} - \mathrm{E}X_i^2\mathbb{1}_{|X_i|>T}\right| \leqslant \\ &\leqslant \mathrm{E}\left|X_i^2\mathbb{1}_{|X_i|>\hat{T}} - X_i^2\mathbb{1}_{|X_i|>T}\right| + \\ &+ \mathrm{E}\left|X_i^2\mathbb{1}_{|X_i|>T} - \mathrm{E}X_i^2\mathbb{1}_{|X_i|>T}\right|. \qquad (11)\end{aligned}$$

Рассмотрим первое слагаемое. Предположим, что $a_i > 0$ (случай $a_i \leqslant 0$ рассматривается аналогично). В силу независимости X_i и $\hat{T}$ при достаточно больших N (таких, что $T - a_i > 0$)

$$\begin{aligned}&\mathrm{E}\left|X_i^2\mathbb{1}_{|X_i|>\hat{T}} - X_i^2\mathbb{1}_{|X_i|>T}\right| = \\ &= \mathrm{E}X_i^2\left|\mathbb{1}_{\hat{T}\geqslant|X_i|>T} + \mathbb{1}_{T\geqslant|X_i|>\hat{T}}\right| \leqslant \\ &\leqslant \mathrm{E}\hat{T}^2\mathbb{1}_{\hat{T}\geqslant|X_i|>T} + \mathrm{E}T^2\mathbb{1}_{T\geqslant|X_i|>\hat{T}} \leqslant \mathrm{E}\hat{T}^2\mathbb{1}_{|X_i|>T} + \\ &+ \frac{1}{\sqrt{2\pi\sigma^2}}\left(\mathrm{E}T^2|\hat{T}-T|\mathbb{1}_{\hat{T}\leqslant a_i} + \right. \\ &\left. + \mathrm{E}T^2 e^{-(\hat{T}-a_i)^2/(2\sigma^2)}|\hat{T}-T|\mathbb{1}_{\hat{T}>a_i}\right).\end{aligned}$$

Учитывая (8) и (9), так же как в работе [8], можно убедиться, что для некоторой константы C_1 справедливо

$$\mathrm{E}T^2 e^{-(\hat{T}-a_i)^2/(2\sigma^2)}|\hat{T}-T|\mathbb{1}_{\hat{T}>a_i} \leqslant \frac{C_1(\ln N)^{3/2}}{N}. \qquad (12)$$

Далее, начиная с некоторого N

$$\begin{aligned}\mathrm{E}T^2|\hat{T}-T|\mathbb{1}_{\hat{T}\leqslant a_i} &\leqslant 2\sqrt{2}\sigma^3(\ln N)^{3/2}\mathrm{E}\mathbb{1}_{\hat{T}\leqslant a_i} \leqslant \\ &\leqslant \frac{C_2(\ln N)^{3/2}}{N} \qquad (13)\end{aligned}$$

с некоторой константой C_2.

Наконец, для некоторой константы C_3, пользуясь (8), (9) и оценкой вероятности больших отклонений для нормального распределения [18], получаем:

$$\mathrm{E}\hat{T}^2\mathbb{1}_{|X_i|>T} = \mathrm{E}\hat{T}^2\mathrm{E}\mathbb{1}_{|X_i|>T} \leqslant \frac{C_3(\ln N)^{1/2}}{N}. \qquad (14)$$

Второе слагаемое в (11) оценивается точно так же, как в работе [8]. Для некоторой константы C_4 справедливо

$$\mathrm{E}\left|X_i^2\mathbb{1}_{|X_i|>T} - \mathrm{E}X_i^2\mathbb{1}_{|X_i|>T}\right| \leqslant \frac{C_4(\ln N)^{1/2}}{N}. \qquad (15)$$

Объединяя (11)–(15), получаем, что существует такая константа C^*, что первая сумма в (10) не превосходит $C^*(\ln N)^{3/2}$. Вторая и третья суммы в (10) оцениваются аналогично. Таким образом, существует такая константа C^{**}, что

$$\mathrm{E}\,|W_N| \leqslant C^{**}(\ln N)^{3/2}. \qquad (16)$$

Далее, для произвольного $\varepsilon > 0$ справедливо

$$\begin{aligned}&\sup_{x\in\mathbf{R}}\left|\mathrm{P}\left(\frac{\widehat{R}_N(f) - R_N(f)}{\sigma^2\sqrt{2N}} < x\right) - \Phi_\Sigma(x)\right| = \\ &= \sup_{x\in\mathbf{R}}\left|\mathrm{P}\left(\frac{V_N + U_N + W_N + Z_N}{\sigma^2\sqrt{2N}} < x\right) - \Phi_\Sigma(x)\right| \leqslant \\ &\leqslant \sup_{x\in\mathbf{R}}\left|\mathrm{P}\left(\frac{V_N + U_N}{\sigma^2\sqrt{2N}} < x\right) - \Phi_\Sigma(x)\right| + \\ &+ \frac{\varepsilon}{\sqrt{2\pi\Sigma}} + \mathrm{P}\left(|W_N + Z_N| > \varepsilon\sigma^2\sqrt{2N}\right).\end{aligned}$$

Выберем $\varepsilon = (\ln N)^{1/2+1/(4(\alpha+1))}N^{1/(4(\alpha+1)-1/4)}$. Тогда, учитывая оценку для $\mathrm{E}\,|Z_N|$ и (16), получаем, что для некоторой константы B_3 справедливо

$$\sup_{x\in\mathbf{R}}\left|\mathrm{P}\left(\frac{\widehat{R}_N(f)-R_N(f)}{\sigma^2\sqrt{2N}}<x\right)-\Phi_\Sigma(x)\right|\leqslant$$
$$\leqslant\sup_{x\in\mathbf{R}}\left|\mathrm{P}\left(\frac{V_N+U_N}{\sigma^2\sqrt{2N}}<x\right)-\Phi_\Sigma(x)\right|+$$
$$+\frac{B_3(\ln N)^{1/2+1/(4(\alpha+1))}}{N^{1/4-1/(4(\alpha+1))}}. \quad (17)$$

Так как V_N и U_N независимы, имеем [19]:

$$\sup_{x\in\mathbf{R}}\left|\mathrm{P}\left(\frac{V_N+U_N}{\sigma^2\sqrt{2N}}<x\right)-\Phi_\Sigma(x)\right|\leqslant$$
$$\leqslant\sup_{x\in\mathbf{R}}\left|\mathrm{P}\left(\frac{V_N}{\sigma^2\sqrt{2N}}<x\right)-\Phi_{\Sigma'}(x)\right|+$$
$$+\sup_{x\in\mathbf{R}}\left|\mathrm{P}\left(\frac{U_N}{\sigma^2\sqrt{2N}}<x\right)-\Phi(x)\right|, \quad (18)$$

где $\Phi_{\Sigma'}(x)$ — функция распределения нормального закона с нулевым средним и дисперсией $\Sigma'=2[4\xi_{3/4}\phi(\xi_{3/4})]^{-2}$; $\Phi(x)$ — функция распределения стандартного нормального закона.

Учитывая результаты работ [12, 17], можно показать, что

$$\sup_{x\in\mathbf{R}}\left|\mathrm{P}\left(\frac{V_N}{\sigma^2\sqrt{2N}}<x\right)-\Phi_{\Sigma'}(x)\right|\leqslant\frac{B_4(\ln N)^{3/4}}{N^{1/4}} \quad (19)$$

для некоторой константы B_4.

Для второго слагаемого, поступая так же, как в работе [8], получаем:

$$\sup_{x\in\mathbf{R}}\left|\mathrm{P}\left(\frac{U_N}{\sigma^2\sqrt{2N}}<x\right)-\Phi(x)\right|\leqslant$$
$$\leqslant\frac{B_5}{N^{1/2}}+\frac{B_6}{N^{1-1/(2(\alpha+1))}} \quad (20)$$

с некоторыми константами B_5 и B_6. Объединяя (17)–(20), получаем (6). Теорема доказана.

В (6) разность $\widehat{R}_N(f)-R_N(f)$ нормируется величиной, зависящей от σ^2. Однако, поскольку в (4) в $\widehat{R}_N(f)$ вместо σ^2 подставляется $\hat{\sigma}^2$, естественнее подставить $\hat{\sigma}^2$ и в эту нормировку. При этом из доказанной теоремы можно получить следующее следствие.

Следствие. Если при выполнении условий теоремы 2 вместо σ^2 в (6) подставить $\hat{\sigma}^2$, то для константы C_0 из теоремы 2 и некоторой константы B_0 справедливо

$$\sup_{x\in\mathbf{R}}\left|\mathrm{P}\left(\frac{\widehat{R}_N(f)-R_N(f)}{\hat{\sigma}^2\sqrt{2N}}<x\right)-\Phi_\Sigma(x)\right|\leqslant$$
$$\leqslant\frac{C_0(\ln N)^{1/2+1/(4(\alpha+1))}}{N^{1/4-1/(4(\alpha+1))}}+\frac{B_0(\ln N)^{1/2}}{N^{1/2}}. \quad (21)$$

Доказательство. Для достаточно малых $\varepsilon>0$

$$\sup_{x\in\mathbf{R}}\left|\mathrm{P}\left(\frac{\widehat{R}_N(f)-R_N(f)}{\hat{\sigma}^2\sqrt{2N}}<x\right)-\Phi_\Sigma(x)\right|=$$
$$=\sup_{x\in\mathbf{R}}\left|\mathrm{P}\left(\frac{\widehat{R}_N(f)-R_N(f)}{\sigma^2\sqrt{2N}}\,\frac{\sigma^2}{\hat{\sigma}^2}<x\right)-\Phi_\Sigma(x)\right|\leqslant$$
$$\leqslant\sup_{x\in\mathbf{R}}\left|\mathrm{P}\left(\frac{\widehat{R}_N(f)-R_N(f)}{\sigma^2\sqrt{2N}}<x\right)-\Phi_\Sigma(x)\right|+$$
$$+\mathrm{P}\left(\left|\frac{\sigma}{\hat{\sigma}}-1\right|>\varepsilon\right)+\frac{3\varepsilon}{\sqrt{2\pi e}}. \quad (22)$$

Далее, используя экспоненциальные неравенства для квантилей и абсолютного медианного отклонения при оценке второго слагаемого в (22) [12, 16], можно показать, что найдутся такие константы B_0^* и B_0, что если положить $\varepsilon=B_0^*(\ln N)^{1/2}N^{-1/2}$, то

$$\mathrm{P}\left(\left|\frac{\sigma}{\hat{\sigma}}-1\right|>\varepsilon\right)+\frac{3\varepsilon}{\sqrt{2\pi e}}\leqslant\frac{B_0(\ln N)^{1/2}}{N^{1/2}}. \quad (23)$$

Объединяя (22), (23) и (6), получаем (21).

Замечание 1. В утверждениях этого раздела требуется равномерная регулярность по Липшицу, когда дисперсия оценивается по независимой выборке. Но это требование можно ослабить, позволив функции быть разрывной в конечном числе точек, если потребовать, чтобы вейвлет-функция имела компактный носитель. При этом порядок оценок в теореме 2 и ее следствии не изменится.

4 Оценивание дисперсии по выборке сигнала

Если дисперсия оценивается по выборке сигнала и функция f удовлетворяет требуемым условиям регулярности, то в силу (2) обычно ее оценивают по половине всех вейвлет-коэффициентов для $j=J-1$ (напомним, что $N=2^J$), так как эти коэффициенты фактически содержат только шум. Для доказательства утверждений этого пункта будем использовать две оценки σ, каждая из которых построена с использованием одной из формул (5) по половине вейвлет-коэффициентов из указанного множества, т. е. по четверти всех вейвлет-коэффициентов. Обозначим эти оценки через $\hat{\sigma}_1$ и $\hat{\sigma}_2$. При пороговой обработке для построения порога $\hat{T}$ будем использовать $\hat{\sigma}_1^2$ для тех наблюдений X_i,

которые не зависят от $\hat{\sigma}_1^2$, и $\hat{\sigma}_2^2$ для тех наблюдений X_i, которые не зависят от $\hat{\sigma}_2^2$. Для наблюдений, не зависящих ни от $\hat{\sigma}_1^2$, ни от $\hat{\sigma}_2^2$, будем использовать одну из этих оценок так, чтобы каждая из них использовалась одинаковое число раз. Таким образом, многие рассуждения, изложенные в теореме 2, останутся справедливыми.

Используя [8–10] и результаты работы [6], можно показать, что если f регулярна по Липшицу с параметром $\gamma = 1/2 + \alpha$, то при использовании интерквартильного размаха

$$\left.\begin{aligned} &\mathrm{E}\hat{\sigma}_k^2 = \sigma^2 + O\left(\frac{1}{N^{3/4}}\right) + O\left(\frac{1}{N^{1/2+\alpha}}\right);\\ &\mathrm{D}\hat{\sigma}_k = O\left(\frac{1}{N}\right), \quad k = 1,2\,, \end{aligned}\right\} \tag{24}$$

а при использовании абсолютного медианного отклонения

$$\left.\begin{aligned} &\mathrm{E}\hat{\sigma}_k^2 = \sigma^2 + O\left(\frac{1}{N^{1/2}}\right);\\ &\mathrm{D}\hat{\sigma}_k = O\left(\frac{1}{N}\right), \quad k = 1,2\,. \end{aligned}\right\} \tag{25}$$

Справедлива следующая теорема.

Теорема 3. *Пусть f задана на отрезке $[0,1]$ и является равномерно регулярной по Липшицу с параметром $\gamma = 1/2 + \alpha$ ($\alpha > 0$), и пусть оценка σ^2 строится по выборке сигнала указанным выше способом. Тогда существует такая константа $\tilde{C}_0$ (зависящая от α, A и σ), что*

$$\sup_{x\in\mathbf{R}}\left|\mathrm{P}\left(\frac{\widehat{R}_N(f) - R_N(f)}{\sigma^2\sqrt{2N}} < x\right) - \Phi_\Upsilon(x)\right| \leqslant \leqslant \frac{\tilde{C}_0(\ln N)^{1/2+1/(4(\alpha+1))})}{N^{1/4-1/(4(\alpha+1))}}, \tag{26}$$

где $\Phi_\Upsilon(x)$ — функция распределения нормального закона с нулевым средним и дисперсией $\Upsilon = [2\xi_{3/4}\phi(\xi_{3/4})]^{-2} - 1$.

Замечание 2. Дисперсия предельного закона в (26) отличается от дисперсии предельного закона в теореме 2.

Доказательство. Так же, как в теореме 2, запишем разность $\widehat{R}_N(f) - R_N(f)$ в виде $\widehat{R}_N(f) - R_N(f) = V_N + U_N + W_N + Z_N$. Рассмотрим сумму $V_N + U_N$:

$$V_N + U_N = \sum_{i\in I_1}\left(X_i^2 - \mathrm{E}X_i^2\right) + N\left(\sigma^2 - \frac{1}{2}\left(\hat{\sigma_1^2} + \hat{\sigma_2^2}\right)\right).$$

Заметим, что оценки $\hat{\sigma}_1$ и $\hat{\sigma}_2$ строятся по слагаемым, индексы которых содержатся в I_1. Обозначим множество этих индексов через I_1'. Таким образом, имеем:

$$V_N + U_N = \sum_{i\in I_1\setminus I_1'}\left(X_i^2 - \mathrm{E}X_i^2\right) + \sum_{i\in I_1'}\left(X_i^2 - \mathrm{E}X_i^2\right) + N\left(\sigma^2 - \frac{1}{2}\left(\hat{\sigma_1^2} + \hat{\sigma_2^2}\right)\right).$$

Пусть

$$U_N' = \sum_{i\in I_1\setminus I_1'}\left(X_i^2 - \mathrm{E}X_i^2\right);$$

$$V_N' = \sum_{i\in I_1'}\left(X_i^2 - \mathrm{E}X_i^2\right) + N\left(\sigma^2 - \frac{1}{2}\left(\hat{\sigma_1^2} + \hat{\sigma_2^2}\right)\right).$$

Так же, как в теореме 2, убеждаемся, что для некоторых констант $\tilde{C}_1$ и $\tilde{C}_2$

$$\sup_{x\in\mathbf{R}}\left|\mathrm{P}\left(\frac{U_N'}{\sigma^2\sqrt{2N}} < x\right) - \Phi_{1/2}(x)\right| \leqslant \leqslant \frac{\tilde{C}_1}{N^{1/2}} + \frac{\tilde{C}_2}{N^{1-1/(2(\alpha+1))}}, \tag{27}$$

где $\Phi_{1/2}(x)$ — функция распределения нормального закона с нулевым средним и дисперсией, равной 1/2. Далее, используя разложение Бахадура [14] и результаты работ [6, 12, 16, 17, 20], а также учитывая (2), можно показать, что для некоторых констант $\tilde{C}_3$ и $\tilde{C}_4$

$$\sup_{x\in\mathbf{R}}\left|\mathrm{P}\left(\frac{V_N'}{\sigma^2\sqrt{2N}} < x\right) - \Phi_{\Upsilon'}(x)\right| \leqslant \leqslant \frac{\tilde{C}_3(\ln N)^{3/4}}{N^{1/4}} + \frac{\tilde{C}_4}{N^{\alpha/2}}, \tag{28}$$

где $\Phi_{\Upsilon'}(x)$ — функция распределения нормального закона с нулевым средним и дисперсией

$$\Upsilon' = [2\xi_{3/4}\phi(\xi_{3/4})]^{-2} - \frac{3}{2}.$$

Наконец, учитывая соотношения (24) и (25), можно оценить сумму $\mathrm{E}\,|W_N| + \mathrm{E}\,|Z_N|$ аналогично тому, как это было сделано в теореме 2. Используя неравенства, аналогичные (17) и (18), и учитывая (27) и (28), получаем (26). Теорема доказана.

Из теоремы 3 можно сделать такое же следствие, как и из теоремы 2.

Следствие. Если при выполнении условий теоремы 3 вместо σ^2 в (26) подставить $\hat{\sigma}^2 = (\hat{\sigma}_1^2 + \hat{\sigma}_2^2)/2$,

то для константы $\tilde{C}_0$ из теоремы 3 и некоторой константы $\tilde{B}_0$ справедливо

$$\sup_{x\in\mathbf{R}}\left|\mathrm{P}\left(\frac{\widehat{R}_N(f)-R_N(f)}{\hat{\sigma}^2\sqrt{2N}}<x\right)-\Phi_\Upsilon(x)\right|\leqslant$$
$$\leqslant\frac{\tilde{C}_0(\ln N)^{1/2+1/(4(\alpha+1))}}{N^{1/4-1/(4(\alpha+1))}}+\frac{\tilde{B}_0(\ln N)^{1/2}}{N^{1/2}}. \quad (29)$$

Доказательство неравенства (29) аналогично доказательству следствия из теоремы 2.

Литература

1. *Donoho D., Johnstone I. M.* Ideal spatial adaptation via wavelet shrinkage // Biometrika, 1994. Vol. 81. No. 3. P. 425–455.
2. *Donoho D., Johnstone I. M.* Adapting to unknown smoothness via wavelet shrinkage // J. Amer. Stat. Assoc., 1995. Vol. 90. P. 1200–1224.
3. *Donoho D., Johnstone I. M., Kerkyacharian G., Picard D.* Wavelet shrinkage: Asymptopia? // J. R. Statist. Soc. Ser. B, 1995. Vol. 57. No. 2. P. 301–369.
4. *Marron J. S., Adak S., Johnstone I. M., Neumann M. H., Patil P.* Exact risk analysis of wavelet regression // J. Comput. Graph. Stat., 1998. Vol. 7. P. 278–309.
5. *Antoniadis A., Fan J.* Regularization of wavelet approximations // J. Amer. Statist. Assoc., 2001. Vol. 96. No. 455. P. 939–967.
6. *Маркин А. В.* Предельное распределение оценки риска при пороговой обработке вейвлет-коэффициентов // Информатика и её применения, 2009. Т. 3. Вып. 4. С. 57–63.
7. *Маркин А. В., Шестаков О. В.* О состоятельности оценки риска при пороговой обработке вейвлет-коэффициентов // Вестн. Моск. ун-та. Сер. 15. Вычисл. матем. и киберн., 2010. № 1. С. 26–34.
8. *Шестаков О. В.* Аппроксимация распределения оценки риска пороговой обработки вейвлет-коэффициентов нормальным распределением при использовании выборочной дисперсии // Информатика и её применения, 2010. Т. 4. Вып. 4. С. 73–81.
9. *Добеши И.* Десять лекций по вейвлетам. — Ижевск: Регулярная и хаотическая динамика, 2001.
10. *Mallat S.* A wavelet tour of signal processing. — New York: Academic Press, 1999.
11. *Abramovich F., Silverman B. W.* Wavelet decomposition approaches to statistical inverse problems // Biometrika, 1998. Vol. 85. No. 1. P. 115–129.
12. *Serfling R.* Approximation theorems of mathematical statistics. — New York: John Wiley and Sons, 1980.
13. *Hall P., Welsh A. H.* Limits theorems for median deviation // Ann. Inst. Stat. Math., 1985. Vol. 37. No. 1. P. 27–36.
14. *Bahadur R. R.* A note on quantiles in large samples // Ann. Statist., 1966. Vol. 37. No. 3. P. 577–580.
15. *Duttweiler D. L.* The mean-square error of Bahadur's order-statistic approximation // Ann. Statist., 1973. Vol. 1. No. 3. P. 446–453.
16. *Serfling R., Mazumder S.* Exponential probability inequality and convergence results for the median absolute deviation and its modifications // Stat. Prob. Lett., 2009. Vol. 79. No. 16. P. 1767–1773.
17. *Mazumder S., Serfling R.* Bahadur representations for the median absolute deviation and its modifications // Stat. Prob. Lett., 2009. Vol. 79. No. 16. P. 1774–1783.
18. *Феллер В.* Введение в теорию вероятностей и ее приложения. — М.: Мир, 1984.
19. *Senatov V. V.* Normal approximation: New results, methods, and problems. — Utrecht: VSP, 1998.
20. *DasGupta A.* Asymptotic values and expansions for the correlation between different measures of spread // J. Stat. Planning Inference, 2006. Vol. 136. No. 7. P. 2197–2212.

РЕЦЕНЗИИ

НОВАЯ КНИГА И. Н. СИНИЦЫНА, А. С. ШАЛАМОВА «ЛЕКЦИИ ПО ТЕОРИИ ИНТЕГРИРОВАННОЙ ЛОГИСТИЧЕСКОЙ ПОДДЕРЖКИ» (М.: ТОРУС ПРЕСС, 2012. 624 с.)

Д.ф.-м.н., профессор С. Я. Шоргин

В книге представлено системное изложение теоретических основ одного из новейших направлений в области экономики послепродажного обслуживания изделий наукоемкой продукции (ИНП) длительного пользования — интегрированной логистической поддержки (ИЛП).

Приведены также результаты новых работ, выполненных в Институте проблем информатики Российской академии наук в рамках научного направления «Информационные технологии и анализ сложных систем».

Излагаемые в книге научные подходы позволяют кардинально реформировать существующие системы производства и эксплуатации ИНП путем создания и внедрения методов рационального и оптимального управления процессами расходования временны́х, материальных, трудовых и других ресурсов на всех стадиях жизненного цикла изделий (ЖЦИ) по критериям экономической целесообразности и эффективности.

В книге приведен краткий обзор причин возникновения и развития CALS-методологии как основы современных международных стандартов по созданию и функционированию глобальных информационно-коммуникационных систем, ее ключевых возможностей и эффективности результатов ее использования. Авторы предлагают ряд научных обоснований для разработки единой теории проектирования и управления систем ИЛП для полноценного использования преимуществ существующей методологии, определяют общую структурную схему комплексной системы «ИНП-СППО» и необходимость разработки для ее описания гибридных стохастических моделей.

Книга состоит из пяти частей, где последовательно излагается материал по каждой из следующих тем: «Интегрированная логистическая поддержка», «Теория гибридных стохастических систем и компьютерная поддержка исследований и разработок», «Основы математического моделирования, анализа и синтеза систем послепродажного обслуживания», «Определение и анализ показателей экспортного потенциала ИНП при проектировании», «Задачи управления поддержкой послепродажного обслуживания», а также «Моделирование инвестиционных процессов ИЛП в условиях неравновесных финансовых рынков».

В конце каждой главы приведены выводы и даны вопросы и задания для самоконтроля. В приложениях содержатся основные определения по программам работ по анализу ИЛП, логистическим базам данных и компьютерным решениям, эквивалентной статистической линеаризации нелинейных преобразований ИЛП, справочный материал, а также развернутые уравнения для вероятностных характеристик.

Книга заинтересует широкий круг специалистов и может быть использована научными проектными организациями в сфере промышленного производства ИНП. Большое количество иллюстраций, примеров и вопросов, обращенных к читателю, позволяет использовать книгу также в качестве учебного пособия для студентов и аспирантов машиностроительных, транспортных и других специальностей, а также для самостоятельного изучения.

Книга представляет несомненный интерес для специалистов и студентов в области прикладной математики и информатики.

ABSTRACTS

ALGORITHM FOR COMPUTATION CHARACTERISTICS OF TELECOMMUNICATION NETWORK WITH RESUBMITS AND INCOMPLETE CIRCUIT BUFFER MANAGEMENT MODEL

Ya. M. Agalarov

IPI RAN, agglar@ya.ru

Telecommunication network with SMQMA (Sharing with Maximum Queue Length and Minimum Allocation) nodes buffer management scheme and the ability to repeat the transmission from the source and transit nodes are discussed. The computation algorithm for average network characteristics such as probability of nodes blocking, total load in lines, average number of packets in nodes and network, average number of packets, waiting for repeat in sources is suggested. Proofs of statements concerning the properties of the algorithm and the results of computational experiments are presented.

Keywords: telecommunication networks; repeat of transmission; mechanisms of buffers management

ANALYSIS AND OPTIMIZATION PROBLEMS FOR SOME USERS ACTITITY MODEL. PART 3. EXTERNAL RESOURCES OPTIMIZATION

A. V. Bosov

IPI RAN, AVBosov@ipiran.ru

This paper completes the mathematical model describing the activity of users and optimization problems for distribution of internal computational resources proposed by the author earlier. The optimization problem for the distribution of external resources is formulated and solved. Suboptimal optimization algorithms are proposed.

Keywords: information systems; database management system; stochastic observation system; quadratic criterion

ON STABILITY OF NORMAL LOCATION MIXTURES WITH RESPECT TO VARIATIONS IN MIXING DISTRIBUTION

A. K. Gorshenin

IPI RAN, agorshenin@ipiran.ru

The stability of finite location mixtures of normal distributions with respect to parameter variations of the mixing distribution is studied. The results are stated for the models of addition and splitting of components which are used to test statistical hypotheses about the number of mixture components.

Keywords: location mixtures of normal distributions; Lévy metric

GEOSPATIAL INFORMATION PROCESSING ON THE BASE OF GIS REPOSITORY

S. K. Dulin[1], I. N. Rozenberg[2], and V. I. Umansky[3]

[1]IPI RAN, s.dulin@ccas.ru

[2]Research & Design Institute for Information Technology, Signalling and Telecommunications on Railway Transport (JSC NIIAS), I.Rozenberg@gismps.ru

[3]"IntechGeoTrans" Close Corporation, umanvi@yandex.ru

Effective integration of the data describing the interacting components is a key to successful management of the entire system of operating objects. Lack of data integration can lead to significant inefficiencies of operational,

tactical, and long-term management strategies. The integrated management system can serve to overcome this inefficiency and improve coordination and cost-effectiveness of solutions. Data Integration is no doubt the most responsible action for the implementation of successful management strategies. An approach to the use of the centralized repository of corporate data which allows to combine spatial and nonspatial data object descriptions is presented. Repository is the environment for data sharing and pooling of data and software products in action.

Keywords: facilities management; data integration; GIS; repository

TECHNIQUE FOR MODELING THE SERVER LOADING IN OPEN CLOUD COMPUTING SYSTEMS

D. V. Zhevnerchuk[1] and A. V. Nikolaev[2]

[1]Tchaikovsky Technological Institute (branch) the Izhevsk State Technical University, drevnigeck@yandex.ru
[2]Tchaikovsky Technological Institute (branch) the Izhevsk State Technical University, elodssa@yandex.ru

Planning resource of server system of cloud computing is a challenging task. Such factors as composition and parameters of the hardware platform, parameters of the system software that manages the execution of applications, the properties of traffic generated by users and determining the modes of operation of applications should be taken into account. Technique of estimating the server loading in open cloud computing systems is proposed. The technique is based on the analysis of processes of user interaction with the software. Reliability of the technique is justified. The results of modeling the load on the server side of management system simulation are presented.

Keywords: cloud computing; imitation modeling; human–computer interaction

RUSSIAN ACADEMY OF SCIENCES LIBRARIES TASKS AND FUNCTIONS IN MODERN CONDITIONS

N. E. Kalenov

Library for Natural Sciences, Russian Academy of Sciences, nek@benran.ru

The problems of scientific libraries activities changes due to the development of network technology, the burgeoning growth of Internet available electronic publications and databases are being considered. It is shown that academic libraries are an integral part of the scientific infrastructure in the modern situation. The traditional tasks such as science information support remain the prerogative of the libraries, but they must be based on new solutions and performed with an extensive use of modern network technologies. New approaches to the traditional tasks are illustrated on the example of Library for Natural Sciences of the Russian Academy of Sciences (LNS RAS) (http://www.benran.ru). In addition to the traditional tasks, academic libraries are ready for new challenges in the area of bibliometrics research, digitization of printed publication materials, etc.

Keywords: academic libraries; automation; informatics; user service; information providing; computer technologies; digital libraries; automation workplaces; LNS RAS

UNIFICATION OF THE RULE-BASED SYSTEM LANGUAGES TO PROVIDE INTEROPERABILITY OF DECLARATIVE PROGRAMS

L. A. Kalinichenko[1] and S. A. Stupnikov[2]

[1]IPI RAN, leonidk@synth.ipi.ac.ru
[2]IPI RAN, ssa@ipi.ac.ru

The W3C standard RIF (Rule Interchange Format) that is provided for the interoperability of various rule based systems by the introduction of the extensible family of unified languages (dialects) oriented on creation of semantic preserving mappings of rule based languages of various systems into the dialects is analyzed. To characterize the motivation for the RIF project, a short survey of development and application of rule based languages and systems

in the areas of knowledge representation, deductive databases, and logical reasoning is made. Various semantics of logical rule based languages that influenced the RIF decisions are also analyzed. Main classes of application cases of the interoperable rule based programs used for development of the requirements for RIF are considered. Finally, the main decisions of the RIF project are overviewed.

Keywords: language unification; language extensibility; logic programming systems; active rule systems; production systems; knowledge representation; deductive databases; logical models of reasoning; stratified semantics; stable model of a logic program; well-founded semantics; RIF dialects; RIF Framework

COGNITIVE STUDY OF AN ASSISTIVE MULTIMODAL USER INTERFACE FOR HANDS-FREE HUMAN–COMPUTER INTERACTION

A. A. Karpov

St. Petersburg Institute for Informatics and Automation, Russian Academy of Sciences (SPIIRAS), karpov@iias.spb.su

The process of research and development of a multimodal user interfaces aimed at hands-free interaction with a personal computer by means of the natural speech input and pointing gestures/movements by user's head is described. The proposed interface uses a low-cost audio-visual equipment for information input and provides an universal access to computer systems both for regular human-operators for contactless (without using hands) personal computer control, and for handicapped users having difficulties with hand motion or even without arms. The methods and results of quantitative evaluation of speech and performance and contactless human–computer interaction are described and comparison with contact-based means of information input is made.

Keywords: multimodal user interface; automatic speech recognition; computer vision; cognitive research

A LOGIC OF BIOGRAPHICAL FACTS

N. A. Markova

IPI RAN, nMarkova@ipiran.ru

A method for formalizing biographical facts in the form of logical formulas is proposed. The method enables to integrate and analyze data obtained from heterogeneous sources and allows to improve the efficiency of the reference-informational service of biographical resources.

Keywords: biographical research; information retrieval; formalization; biographical fact

CALCULATION AND OPTIMIZATION OF SOME CHARACTERISTICS OF THE MODEL FOR COMPUTATIONAL COMPLEX

I. V. Pavlov

Bauman Moscow State Technical University, ipavlov@mbstu.ru

The problem of optimal packet sizing while processing large volume information is considered. In the model of a computer system, possible failures or faults of elements during tasks execution are taken into account. An asymptotic solution is obtained for the case of highly reliable components and small packets transfer time.

Keywords: optimal packet size; reliability; failure rate; transfer time

FUZZY VARIABLES AS A TOOL FOR EXPRESSING ERROR CHARACTERISTICS IN DATA PROCESSING

K. K. Semenov

St. Petersburg State Polytechnic University, semenov.k.k@gmail.com

The main application of fuzzy variables theory is to take into account badly formalized information about uncertainty sources. In this paper, the results of another application are presented: to take into account well-formalized information about errors that is classical for metrology. The principal possibility of error characteristics representation as fuzzy variables that is compatible with metrological norms and standards has been shown.

Keywords: fuzzy variables; error characteristics; measurement results

SEMANTIC SEARCH OF NATURAL LANGUAGE INFORMATION ON THE BASIS OF KNOWLEDGE BASE TECHNOLOGY

M. M. Sharnin[1] and I. P. Kuznetsov[2]

[1]IPI RAN, keywen1@mail.com

[2]IPI RAN, igor-kuz@mtu-net.ru

A system for semantic search in the natural language texts is considered. The search is based on the use of a linguistic processor which analyzes the text in natural language and extracts from it the information objects (named entities), their signs, links, and participation in actions. As a result, the processor forms the semantic structures in the Knowledge Base. The structure of quiries is formed by analogy. The semantic search consists in comparison of such structures and finding the unknown objects. In the process, the connections between objects and their participation in the actions are taken into account.

Keywords: semantic search; semantics oriented linguistic processor; knowledge extraction from texts; Knowledge Base

ON THE RATE OF CONVERGENCE TO THE NORMAL LAW OF RISK ESTIMATE FOR WAVELET COEFFICIENTS THRESHOLDING WHEN USING ROBUST VARIANCE ESTIMATES

O. V. Shestakov

Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics,
M. V. Lomonosov Moscow State University; IPI RAN, oshestakov@cs.msu.su

The asymptotic properties of risk estimate for thresholding wavelet coefficients of signal function are analyzed. Some estimates for the rate of convergence to the normal law are obtained.

Keywords: wavelets; thresholding; risk estimate; normal distribution; rate of convergence

Об авторах

Агаларов Явер Мирзабекович (р. 1952) — кандидат технических наук, доцент, ведущий научный сотрудник ИПИ РАН

Босов Алексей Вячеславович (р. 1969) — доктор технических наук, заведующий сектором ИПИ РАН

Горшенин Андрей Константинович (р. 1986) — кандидат физико-математических наук, старший научный сотрудник ИПИ РАН

Дулин Сергей Константинович (р. 1950) — доктор технических наук, профессор, старший научный сотрудник ИПИ РАН

Жевнерчук Дмитрий Валерьевич (р. 1978) — кандидат технических наук, доцент Чайковского технологического института (филиала) Ижевского государственного технического университета

Калёнов Николай Евгеньевич (р. 1945) — доктор технических наук, профессор, директор Библиотеки по естественным наукам РАН

Калиниченко Леонид Андреевич (р. 1937) — доктор физико-математических наук, профессор, заслуженный деятель науки РФ, заведующий лабораторией ИПИ РАН

Карпов Алексей Анатольевич (р. 1978) — кандидат технических наук, старший научный сотрудник Санкт-Петербургского института информатики и автоматизации РАН

Кузнецов Игорь Петрович (р. 1938) — доктор технических наук, профессор, главный научный сотрудник ИПИ РАН

Маркова Наталья Александровна (р. 1950) — кандидат физико-математических наук, ведущий научный сотрудник ИПИ РАН

Николаев Андрей Валерьевич (р. 1985) — кандидат технических наук, старший преподаватель Чайковского технологического института (филиала) Ижевского государственного технического университета

Павлов Игорь Валерианович (р. 1945) — доктор физико-математических наук, профессор МГТУ им. Н. Э. Баумана

Розенберг Игорь Наумович (р. 1965) — доктор технических наук, первый заместитель директора ОАО «Научно-исследовательский и проектно-конструкторский институт информатизации, автоматизации и связи на железнодорожном транспорте» (ОАО «НИИАС»)

Семёнов Константин Константинович (р. 1986) — магистр, доцент Санкт-Петербургского государственного политехнического университета

Ступников Сергей Александрович (р. 1978) — кандидат технических наук, старший научный сотрудник ИПИ РАН

Уманский Владимир Ильич (р. 1954) — кандидат технических наук, генеральный директор ЗАО «ИнтехГеоТранс»

Шарнин Михаил Михайлович (р. 1959) — кандидат технических наук, старший научный сотрудник ИПИ РАН

Шестаков Олег Владимирович (р. 1976) — кандидат физико-математических наук, ассистент кафедры математической статистики факультета вычислительной математики и кибернетики МГУ им. М. В. Ломоносова; старший научный сотрудник ИПИ РАН

About Authors

Agalarov Yaver M. (b. 1952) — Candidate of Science (PhD) in technology, leading scientist, Institute of Informatics Problems, Russian Academy of Sciences

Bosov Alexey V. (b. 1969) — Doctor of Science in technology, Head of Laboratory, Institute of Informatics Problems, Russian Academy of Sciences

Dulin Sergey K. (b. 1950) — Doctor of Science in technology, professor, senior scientist, Institute of Informatics Problems, Russian Academy of Sciences

Gorshenin Andrey K. — (b. 1986) — Candidate of Science (PhD) in physics and mathematics, senior scientist, Institute of Informatics Problems, Russian Academy of Sciences

Kalenov Nikolay E. (b. 1945) — Doctor of Science in technology, professor, Director, Library for Natural Sciences, Russian Academy of Sciences

Kalinichenko Leonid A. (b. 1937) — Doctor of Science in physics and mathematics, professor, Honored scientist of RF, Head of Laboratory, Institute of Informatics Problems, Russian Academy of Sciences

Karpov Alexey A. (b. 1978) — Candidate of Science (PhD) in technology, senior scientist, St. Petersburg Institute for Informatics and Automation, Russian Academy of Sciences

Kuznetsov Igor P. (b. 1938) — Doctor of Science in technology, professor, principal scientist, Institute of Informatics Problems, Russian Academy of Sciences

Markova Natalia A. (b. 1950) — Candidate of Science (PhD) in physics and mathematics, leading scientist, Institute of Informatics Problems, Russian Academy of Sciences

Nikolaev Andrey V. (b. 1985) — Candidate of Science (PhD) in technology, senior lecturer, Tchaikovsky Technological Institute, Branch of the Izhevsk State Technical University

Pavlov Igor V. (b. 1945) — Doctor of Science in physics and mathematics, professor, Bauman Moscow State Technical University

Rozenberg Igor N. (b. 1965) — Doctor of Science in technology, First Deputy Director General, Research & Design Institute for Information Technology, Signalling and Telecommunications on Railway Transport (JSC NIIAS)

Semenov Konstantin K. (b. 1986) — MPhil, associate professor, St. Petersburg State Polytechnical University

Sharnin Mikhail M. (b. 1959) — Candidate of Science (PhD) in technology, senior scientist, Institute of Informatics Problems, Russian Academy of Sciences

Shestakov Oleg V. (b. 1976) — Candidate of Science (PhD) in physics and mathematics, assistant professor, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M.V. Lomonosov Moscow State University; senior scientist, Institute of Informatics Problems, Russian Academy of Sciences

Stupnikov Sergey A. (b. 1978) — Candidate of Science (PhD) in technology, senior scientist, Institute of Informatics Problems, Russian Academy of Sciences

Umansky Vladimir I. (b. 1954) — Candidate of Science (PhD) in technology, Director General, "IntechGeoTrans" Closed Joint Stock Company

Zhevnerchuk Dmitry V. (b. 1978) — Candidate of Science (PhD) in technology, associate professor, Tchaikovsky Technological Institute, Branch of the Izhevsk State Technical University

Правила подготовки рукописей статей для публикации в журнале «Информатика и её применения»

Журнал «Информатика и её применения» публикует теоретические, обзорные и дискуссионные статьи, посвященные научным исследованиям и разработкам в области информатики и ее приложений. Журнал издается на русском языке. По специальному решению редколлегии отдельные статьи, в виде исключения, могут печататься на английском языке. Тематика журнала охватывает следующие направления:

- теоретические основы информатики;
- математические методы исследования сложных систем и процессов;
- информационные системы и сети;
- информационные технологии;
- архитектура и программное обеспечение вычислительных комплексов и сетей.

1. В журнале печатаются результаты, ранее не опубликованные и не предназначенные к одновременной публикации в других изданиях. Публикация не должна нарушать закон об авторских правах. Направляя свою рукопись в редакцию, авторы автоматически передают учредителям и редколлегии неисключительные права на издание данной статьи на русском языке и на ее распространение в России и за рубежом. При этом за авторами сохраняются все права как собственников данной рукописи. В связи с этим авторами должно быть представлено в редакцию письмо в следующей форме: Соглашение о передаче права на публикацию:

   «Мы, нижеподписавшиеся, авторы рукописи « », передаем учредителям и редколлегии журнала «Информатика и её применения» неисключительное право опубликовать данную рукопись статьи на русском языке как в печатной, так и в электронной версиях журнала. Мы подтверждаем, что данная публикация не нарушает авторского права других лиц или организаций. Подписи авторов: (ф. и. о., дата, адрес)».

   Указанное соглашение может быть представлено как в бумажном виде, так и в виде отсканированной копии (с подписями авторов).

   Редколлегия вправе запросить у авторов экспертное заключение о возможности опубликования представленной статьи в открытой печати.
2. Статья подписывается всеми авторами. На отдельном листе представляются данные автора (или всех авторов): фамилия, полные имя и отчество, телефон, факс, e-mail, почтовый адрес. Если работа выполнена несколькими авторами, указывается фамилия одного из них, ответственного за переписку с редакцией.
3. Редакция журнала осуществляет самостоятельную экспертизу присланных статей. Возвращение рукописи на доработку не означает, что статья уже принята к печати. Доработанный вариант с ответом на замечания рецензента необходимо прислать в редакцию.
4. Решение редакционной коллегии о принятии статьи к печати или ее отклонении сообщается авторам. Редколлегия не обязуется направлять рецензию авторам отклоненной статьи.
5. Корректура статей высылается авторам для просмотра. Редакция просит авторов присылать свои замечания в кратчайшие сроки.
6. При подготовке рукописи в MS Word рекомендуется использовать следующие настройки. Параметры страницы: формат — A4; ориентация — книжная; поля (см): внутри — 2,5, снаружи — 1,5, сверху — 2, снизу — 2, от края до нижнего колонтитула — 1,3. Основной текст: стиль — «Обычный»: шрифт Times New Roman, размер 14 пунктов, абзацный отступ — 0,5 см, 1,5 интервала, выравнивание — по ширине. Рекомендуемый объем рукописи — не свыше 25 страниц указанного формата. Ознакомиться с шаблонами, содержащими примеры оформления, можно по адресу в Интернете: http://www.ipiran.ru/journal/template.doc.
7. К рукописи, предоставляемой в 2-х экземплярах, обязательно прилагается электронная версия статьи (как правило, в форматах MS WORD (.doc) или LaTeX (.tex), а также — дополнительно — в формате .pdf) на дискете, лазерном диске или по электронной почте. Сокращения слов, кроме стандартных, не применяются. Все страницы рукописи должны быть пронумерованы.
8. Статья должна содержать следующую информацию на русском и английском языках: название, Ф.И.О. авторов, места работы авторов и их электронные адреса, подробные сведения об авторах, оформленные в соответствии с форматом, определяемым файлами http://www.ipiran.ru/journal/issues/2011_05_01/authors.asp и http://www.ipiran.ru/journal/issues/2011_01_eng/authors.asp, аннотация (не более 100 слов), ключевые слова. Ссылки на литературу в тексте статьи нумеруются (в квадратных скобках) и располагаются в порядке их первого упоминания. В списке литературы не должно быть позиций, на которые нет ссылки в тексте статьи. Все фамилии авторов, заглавия статей, названия книг, конференций и т. п. даются на языке оригинала, если этот язык использует кириллический или латинский алфавит.
9. Присланные в редакцию материалы авторам не возвращаются.
10. При отправке файлов по электронной почте просим придерживаться следующих правил:
    - указывать в поле subject (тема) название журнала и фамилию автора;
    - использовать attach (присоединение);
    - в случае больших объемов информации возможно использование общеизвестных архиваторов (ZIP, RAR);
    - в состав электронной версии статьи должны входить: файл, содержащий текст статьи, и файл(ы), содержащий(е) иллюстрации.
11. Журнал «Информатика и её применения» является некоммерческим изданием. Плата за публикацию с авторов не взимается, гонорар авторам не выплачивается.

Адрес редакции: Москва 119333, ул. Вавилова, д. 44, корп. 2, ИПИ РАН
Тел.: +7 (499) 135-86-92 Факс: +7 (495) 930-45-05 E-mail: rust@ipiran.ru