ITHEA

International Journal INFORMATION TECHNOLOGIES & KNOWLEDGE

2011 Volume 5 Number 3

International Journal

INFORMATION TECHNOLOGIES & KNOWLEDGE

Volume 5 / 2011, Number 3

International Journal “INFORMATION TECHNOLOGIES & KNOWLEDGE” (IJ ITK)
is official publisher of the scientific papers of the members of
the ITHEA International Scientific Society

IJ ITK rules for preparing the manuscripts are compulsory.
The **rules for the papers** for IJ ITK as well as the **subscription fees** are given on *www.foibg.com*.

Responsibility for papers published in IJ ITK belongs to authors.

General Sponsor of IJ ITK is the **Consortium FOI Bulgaria** (www.foibg.com).

International Journal “INFORMATION TECHNOLOGIES & KNOWLEDGE” Vol.5, Number 3, 2011

Edited by the **Institute of Information Theories and Applications FOI ITHEA**, Bulgaria, in collaboration with:
Institute of Mathematics and Informatics, BAS, Bulgaria,
V.M.Glushkov Institute of Cybernetics of NAS, Ukraine,
Universidad Politécnica de Madrid, Spain.

Publisher ITHEA®
Sofia, 1000, P.O.B. 775, Bulgaria. www.ithea.org, www.foibg.com, e-mail: info@foibg.com
Printed in Bulgaria

ISSN 1313-0455 (printed) ISSN 1313-048X (online) ISSN 1313-0501 (CD/DVD)

ЕВКЛИДОВЫ ПРОСТРАНСТВА ЧИСЛОВЫХ ВЕКТОРОВ И МАТРИЦ: КОНСТРУКТИВНЫЕ МЕТОДЫ ОПИСАНИЯ БАЗОВЫХ СТРУКТУР И ИХ ИСПОЛЬЗОВАНИЕ

Владимир Донченко

Аннотация: *Определены и рассмотрены базовые линейные структуры матричных евклидовых пространств, которые естественным образом расширяют возможности представления моделируемого объекта до «матрицы признаков» и позволяют непосредственным образом учесть динамику исследуемого объекта. «Базовость» структур: линейных и нелинейных, для матричного случая, как и для евклидовых пространств числовых векторов, означает их принципиальное значение для использования в прикладных задачах. Как и в векторном случае, фундаментальную роль в конструктивном построении матричных базовых структур: их описания, взаимного перехода и использования, является аппарат сингулярного представления для линейных операторов между евклидовыми пространствами. Ту же роль в матричном случае сохраняет и псевдообращение по Муру - Пенроузу для линейных операторов на матричных аргументах. В работе представлены результаты, касающиеся сингулярного разложения и псевдообращения по Муру-Пенроузу для матрицы линейных операторов между матричными евклидовыми пространствами. Введены в рассмотрение кортежные объекты и операции над ними, обобщающие понятия классических числовых векторов и блочных матриц ленточного характера. Развита теория группирующих операторов, важная в прикладном отношении как инструмент выявления и использования групповых свойств объектов в евклидовых пространствах, как для пространств числовых векторов, так и для матричных евклидовых пространств. Кроме того, группирующие операторы позволяют, в частности, прояснить алгебраическую суть расстояния Махаланобиса.*

Ключевые слова*: Псевдообращение по Муру - Пенроузу, сингулярное представление матрицы, кластеризация, расстояние Махаланобиса.*

ACM Classification Keywords: G.3 Probability and statistics, G.1.6. Numerical analysis: Optimization; G.2.m. Discrete mathematics: miscellaneous.

Вступление

В работе предложена и развита концепция базовых структур матричных евклидовых пространства, которая переносит представление о базовых структурах евклидовых пространств числовых векторов на матричный случай. Как и в векторном случае, к «базовым структурам» предлагается отнести объекты двух основных видов: линейные и нелинейные, которые могут носить множественный или «сингловый», единичный характер. К множественным структурам следует отнести подпространства и гиперплоскости (линейные структуры), а также эллипсоиды (нелинейные структуры квадратического характера). К «сингловым» структурам, относятся линейные операторы и квадратические формы, а также матрицы, которые их представляют. Отметим, что к базовым нелинейным структурам отнесены структуры, носящие квадратичный характер: связанные и определяемые подходящими неотрицательно определёнными (симметричными) квадратичными формами. Важность множественных вариантов структур евклидовых пространств определяется естественным характером их использования. Так, к примеру, подпространства и гиперплоскости могут использоваться как для представления классов группирования объектов, так и для их дискриминации. Что касается эллипсоидов, то они, к примеру, естественным образом появляются, в описании неопределённостей в задачах минимаксного характера. Кроме того, они естественным образом могут представлять классы группируемых объектов через минимальные эллипсоиды группировки. Отметим, что примечательной особенностью эллипсоидов группировки [Донченко, 2010] является то, что они содержат все векторы набора векторов, и являются в определённом смысле «минимальными» для них. В работе на случай матричных евклидовы пространств перенесены конструктивные методы генерации базовых структур в связи с порождающей их совокупностью объектов. Кроме того, получены и приведены формулы взаимного перехода от одних типов структур к другим: от линейных подпространств и гиперплоскостей к подходящим матрицам и наоборот, а также от набора векторов к матрицам квадратичных форм и эллипсоидов группировки. В числе других рассмотрены конструктивные способы порождения подпространств и гиперплоскостей, а также ортогональных проекторов, связанных с указанными объектами. В том же русле конструктивности лежит рассмотрение концепции группирующих операторов и минимальных эллипсоидов группировки. Концепция группирующих операторов позволяет использовать в прикладных исследованиях связь того или иного набора векторов с важным видом нелинейных структур евклидовых пространств: минимальными эллипсоидами группировки. И в линейном, и в нелинейном случае упомянутая конструктивность обеспечивается применением псевдообращения по Муру – Пенроузу (ПдО), а также новыми результатами в этой области, берущими своё начало и опирающимися на фундаментальную работу Н.Ф. Кириченко [Кириченко, 1997]. Для расширения возможности аппарата, предложено развитие теории псевдообращения на матричные отображения между евклидовыми пространствами матриц фиксированных размерностей со следовыми скалярными произведениями. Кроме того, на случай отображений между евклидовыми матричными пространствами перенесена теорема о сингулярном разложении матрицы оператора, доказана теорема свёртки о соотношении между сингулярными разложениями одной и той же матрицы в векторном и матричном случаях.

Следует отметить, также, что важную роль в применении аппарата псевдообращения играет сингулярное представление (его называют также SVD - представлением) произвольной $m \times n$ матрицы в специфической форме записи в виде взвешенной суммы тензорных произведений специального набора пар векторов с прозрачным геометрическим содержанием. При этом матрица рассматривается как матрица линейного оператора между двумя пространствами числовых векторов: R^n и R^m. Это определяет специфику сингулярного разложения и интерпретацию его составляющих.

Заметим, однако, что одна и та же $m \times n$ матрица A задаёт как линейный оператор $A: R^n \to R^m$ между евклидовыми пространствами числовых векторов, так и линейный оператор между матричными евклидовыми пространствами $A: R^{n \times p} \to R^{m \times p}$ (p -произвольное натуральное). Переход к евклидовым пространствам $R^{n \times p}, R^{m \times p}$ матриц со следовыми скалярными произведениями требует учёта специфики этих евклидовых пространств для передачи принципиальных особенностей SVD-разложения для такого рода объектов. В работе приведено SVD-разложение для матриц, рассматриваемых как линейные операторы между пространствами матриц, доказана теорема свёртки, связывающие два варианта SVD-представления матрицы: векторного и матричного.

Важным в описании нелинейных структур являются «минимальные» эллипсоиды группировки и, как и в линейном случае, конструктивные способы их описания. Говоря о «минимальных» эллипсоидах группировки, будем иметь в виду «минимальный» в определённом смысле эллипсоид, включающий заданный набор векторов. Как оказывается, матрицей квадратичной формы для такого эллипсоида является группирующий оператор, конструктивно описываемый средствами псевдообращения. Квадратичная форма, отвечающая группирующему оператору, в искажающем суть такого оператора виде, фигурирует в «статистическом» варианте расстояния Махаланобиса [Mahalanobis,1936] (см. также, например, [McLachlan, Geoffry, 1992]), когда вместо ковариационной матрицы многомерного нормального распределения используется её стандартная оценка на основе выборки векторов $a(1),...,a(n)$.

В заключение отметим, что основные идеи, дух и результаты предлагаемой работы восходят к работам и используют результаты развитой в них теории псевдообращения нашего безвременно ушедшего коллеги, друга и учителя, профессора Н.Ф. Кириченко.

Базовые линейные структуры и связи между ними

В дальнейшем, говоря о евклидовом пространстве R^p, будем иметь в виду множество конечных числовых последовательностей одной и той же длины p, записанных в столбик с покоординатными операциями сложения и умножения на скаляр и суммой покоординатных произведений в качестве скалярного произведения. Именно такой вариант евклидового пространства будем стандартным образом обозначать через R^p, а его элементы – через $a, a^T = (a_1,...,a_p)$. Стандартные ортонормированные базисы, составленные из векторов с единственной единичной компонентой (остальные – нули) на месте с соответствующим номером будут обозначаться для R^m и R^n соответственно через $e(j) \in R^m, j = \overline{1,m}, e_{(i)} \in R^n, i = \overline{i,n}$, а для общего случая - $e_k \in R^p, k = \overline{1,p}$. Оператор A из R^n, в R^m: $A: R^n \to R^m$, задаваемый в ортонормированных базисах $e(j) \in R^m, j = \overline{1,m}, e_{(i)} \in R^n, i = \overline{i,n}$, будем отождествлять с $m \times n$ - матрицей $A = (a_{ij})$ этого оператора в этих базисах. Для матрицы $A = (a_{ij})$ будем использовать также блочное представление по столбцам (столбцовое) и строкам (строчное):

$$A = \begin{pmatrix} a_{(1)}^T \\ \cdots \\ a_{(m)}^T \end{pmatrix} = (a(1)\vdots...\vdots a(n)), a_{(i)} \in R^n, i = \overline{1,m}, a(j) \in R^m, j = \overline{1,n} .$$

Пространство всех $p \times q$ матриц с покоординатным умножением и сложением, а также «следовым» скалярным произведением будем обозначать $R^{p \times q}$. «Следовое» скалярное произведение $(\cdot,\cdot)_{tr}$ двух матриц определяется соотношением

$$(C\cdot, D\cdot)_{tr} = trC^T D = \sum_{i=1}^{q} \sum_{k=1}^{p} c_{ki} d_{ki}, C = (c_{ij}), D = (d_{ij}) \in R^{p \times q} .$$

Линейное подпространство, порождённое системой векторов, $c_k \in R^p, k = \overline{1,K}$, будет обозначаться через $L(c_k, k = \overline{1,K}) \equiv L(c_1...,c_K)$, а линейное подпространство значений линейного оператора $A: R^n \to R^m$ – через L_A, соответственно, для A^T - через L_{A^T}.

То же самое будет относиться к линейным подпространствам матричных евклидовых пространств: через $L(C_k, k = \overline{1,K}) \equiv L(C_1...,C_K) \subseteq R^{p \times q}, C_k \in R^{p \times q}, k = \overline{1,K}$ будем обозначать линейное подпространство в $R^{p \times q}$, порождённое набором матриц $C_k \in R^{p \times q}, k = \overline{1,K}$.

Соответственно, через $\Gamma(c, L) = c + L, \Gamma(C, L) = C + L$ в вектором или матричном случае будем обозначать гиперплоскость в соответствующем евклидовом пространстве.

В работе [Донченко, 2010] представлены основные формулы, устанавливающие связь между множественными и сингловыми линейными структурами для евклидовы пространств числовых векторов. Отметим, что аналоги этих формул справедливы и для матричных объектов с достаточно очевидной заменой стандартных единичных ортов в R^n или R^m на аналоги, каковыми являются блочные матрицы, составленные из нулевых и единичных матриц подходящей размерности. Заметим также, что блочные матрицы: строчные или столбцовые, - составленные из блоков одинаковой размерности представляют собой, собственно, строчные или столбцовые кортежи с матричными элементами. Явно вводя длину кортежа в его наименование, будем говорить об n - кортежах строках или n -кортежах столбцах с компонентами из соответствующего множества. Так числовой вектор-столбец из R^n является -кортежем с числовыми элементами, а упорядоченный набор $(C_1...,C_n), C_i \in R^{p\times q}, i = \overline{1,n}$ является n -кортежем-строкой с элементами из матричного евклидова пространства $R^{p\times q}$. Возможны операции с разнородными кортежными конструкциями, обобщающие матричные операции. Так, например можно говорить о произведении двух n - кортежей: строчного и стобцового, определяемого через сумму произведений компонент. Смысл операции произведения в каждом конкретном случае определяется природой компонент кортежей. Так для строчного - кортежа $(C_1...,C_n), C_i \in R^{p\times q}, i = \overline{1,n}$ с матричными компонентами и числового столбцового - кортежа $b : b^T = (b_1...,b_n), b_i \in R^1, i = \overline{1,n}$ (числового вектора $b \in R^n$) произведение $(C_1...,C_n)b$ определяется как

$$(C_1...,C_n)b = \sum_{i=1}^{n} C_i b_i \ .$$

Если же числовые компонент у столбцового n -кортежа b заменить матричными:

$$b = \begin{pmatrix} B_1 \\ \cdots \\ B_n \end{pmatrix}, B_i \in R^{q\times p}, i = \overline{1,n},$$

То: $$(C_1...,C_n)b = \sum_{i=1}^{n} C_i B_i \ .$$

Точно таким же образом можно говорить об блочных матрицах с элементами из подходящих евклидовых пространств и умножении слева или справа таких матриц на строчные или столбцовые кортежи.

Заметим, что вместе с кортежными операциями могут быть определены и обычные матричные. Как правило, результат выполнения кортежных и матричных операций в случае их одновременного рассмотрения, является одинаковыми.

В рамках определённых таким образом операций, обобщающих матричные на кортежные, можно, например, утверждать, что справедливо следующее соотношение, являющееся обобщением теоремы о представлении произведения матриц суммой тензорных произведений компонент сомножителе [Донченко, 2010].

1. *«Тензорное произведение» строчного и столбцового кортежей*. Для произведения матричных - кортежей B, C : строчного и столбцового

$$B = (B(1) \vdots ... \vdots B(r)), B(j) \in R^{p\times q}, j = \overline{1,r}, C^T = \begin{pmatrix} C_{(1)}^T \\ \cdots \\ C_{(r)}^T \end{pmatrix}, C_{(i)} \in R^{q\times k}, i = \overline{1,r} \qquad (1)$$

а также для диагональной матрицы $\Lambda = diag(\lambda_1,...,\lambda_r), \lambda_i \in R^1, i = \overline{1,r}$ справедливо соотношение

$$B\Lambda C = \sum_{i=1}^{r} \lambda_i B(i) C_{(i)}^T \ .$$

В упомянутой выше работе [Донченко, 2010] приведен вариант SVD-представления матрицы оператора между двумя евклидовыми пространствами числовых векторов в форме, допускающей его прямое перенесение на матрицы произвольной размерности, рассматриваемые как матрицы линейных отображений между матричными евклидовыми пространствами. Этот вариант представления приведен ниже под номером 2.

2. *Вариант SVD- представление произвольной* $m \times n$ *матрицы*. Для произвольной $A \in R^{m \times n}$ ранга $r \le \min(m,n)$, рассматриваемой как матрица линейного отображения между евклидовыми пространствами числовых векторов: $A : R^n \to R^m$, справедливо следующее представление матрицы в виде взвешенной суммы тензорных произведений векторов

$$Ax = \sum_{i=1}^{r} \lambda_i u_i (v_i, x)_{R^n}, x \in R^n \tag{2}$$

где

- $\lambda_1^2 \ge ... \ge \lambda_r^2 > 0$ общий набор ненулевых собственных чисел матриц $AA^T, A^T A$;
- $u_i \in R^m, i = \overline{i,r}$ - ортонормированный набор собственных векторов матрицы AA^T, отвечающих ненулевым собственным числам: $AA^T u_i = \lambda_i^2, \lambda_i^2 > 0, i = \overline{1,r}, u_i^T u_j = \delta_{ij}, i \ne j$;
- $v_i \in R^n, i = \overline{1,r}$ -- ортонормированный набор собственных векторов матрицы $A^T A$, отвечающих ненулевым собственным числам: $A^T A v_i = \lambda_i^2 v_i, \lambda_i^2 > 0, i = \overline{1,r}, v_i^T v_j = \delta_{ij}, i \ne j$.

3. *Комментарий к п.2.* Собственно, действие оператора A на вектор $x \in R^n$ в варианте (2) SVD-разложения, представляет собой перенос разложения вектора по ортогональному базису или его части $v_i \in R^n, i = \overline{1,r}$ с координатами $x_i = v_i^T x = (v_i, x)_{R^n}, i = \overline{1,r}$, в одном пространстве в разложение по ортогональному базису $u_i \in R^m, i = \overline{1,r}$, или - его части, в другом пространстве с умножением координат на положительные числа, соответственно, $\lambda_i \in R^m, i = \overline{1,r}$.

Ортогональные проекторы (в дальнейшем ОП) являются важной составляющей аппарата конструктивного описания и использования линейных структур в матричном варианте евклидовых пространств, как в случае евклидовых пространств числовых векторов. И в этом, более общем варианте евклидовых пространств, общей, основой их эффективного использования, является псевдообращение по Муру Пенроузу. Напомним, что в определении по Муру [Moore, 1920] прямо связывается с двумя основными линейными подпространствами линейного оператора: его множеством значений и его ядром. Ортогональных проекторов, как и возможности их конструктивного построения в связи с линейными подпространствами, является наличие двух эквивалентных определений таких проекторов, а также возможность их описание через псевдообращение. Для матрицы, рассматриваемой как линейный оператор между матричными евклидовыми пространствами, приведённое выше соображение остаётся в полной мере справедливым.

Матрицы как линейные операторы между матричными евклидовыми пространствами

4. *Матрица* $A \in R^{m \times n}$ *как линейный оператор из матричного евклидового пространства* $R^{n \times p}$ *в* $R^{m \times p}$. Произвольная $m \times n$ матрица A может рассматриваться как матрица линейного оператора между двумя евклидовыми пространствами матриц $R^{n \times p}$ и $R^{m \times p}$ со следовым скалярным произведением. Этот оператор описывается стандартным образом: для произвольной матрицы $X \in R^{n \times p}$ результатом действия оператора является матрица $AX \in R^{m \times p}$.

5. *Матричное SVD-разложение* $m \times n$ *матрицы* $A \in R^{m \times n}$ *(M-SVD).*

Теорема 1. Справедливо следующее представление матрицы в виде суммы, реализующий перенос разложения по ортонормированным собственным элементам (собственным матрицам) матрицы $A^T A$ на ортонормированные собственные элементы матрицы $A^T A$

$$A \Rightarrow X \sum_{i=1}^{r}\sum_{k=1}^{p} \lambda_i u_i e_k^T (v_i e_k^T, X)_{tr} \quad (3)$$

где

- $\lambda_1^2 \geq ... \geq \lambda_r^2 > 0$ общий набор ненулевых собственных чисел матриц $AA^T, A^T A$;
- $u_i e_k^T \in R^{m\times p}, i = \overline{1,r}, k = \overline{1,p}$ - ортонормированный набор собственных $m \times p$ матриц матрицы AA^T, отвечающих ненулевым собственным числам: $AA^T u_i e_k^T = \lambda_i^2 u_i e_k^T, \lambda_i^2 > 0, i = \overline{1,r}, k = \overline{1,p}, (u_i e_k^T, u_j e_l^T)_{tr} = \delta_{ij}\delta_{kl}, i \neq j, k \neq l$;
- $v_i e_k^T \in R^{n\times p}, i = \overline{1,r}, k = \overline{1,p}$ - ортонормированный набор собственных $n \times p$ матриц матрицы $A^T A$, отвечающих ненулевым собственным числам: $A^T A v_i e_k^T = \lambda_i^2 v_i e_k^T, \lambda_i^2 > 0, i = \overline{1,r}, k = \overline{1,p}, (v_i e_k^T, v_j e_l^T)_{tr} = \delta_{ij}\delta_{kl}, i \neq j, k \neq l$.

6. Совпадение векторного и матричного варианта SVD: теорема свёртки.

Теорема 2. Векторный и матричный вариант SVD-разложения одной и той же $m \times n$ матрицы $A \in R^{m\times n}$ совпадают между собой: для произвольной матрицы $X \in R^{n\times p}$

$$\left(\sum_{i=1}^{r} \lambda_i u_i v_i^T\right) X = \sum_{i=1}^{r}\sum_{k=1}^{p} \lambda_i u_i e_k^T (v_i e_k^T, X)_{tr} \quad (4)$$

Как отмечалось выше, принципиальную роль в описании базовых структур евклидовых пространств играет псевдообращение по Муру – Пенроузу [Moore,1920], [Penrose, 1955] (в дальнейшем ПдО) как одноместной операции A^+ над матрицами произвольной размерности

С учётом п.5 необходимо различать ПдО для матрицы, как векторного отображения, и ПдО для той же матрицы, как матричного отображения. Первое, собственно, классическое по Муру-Пенроузу, будем обозначать как ВПдО, второе – как МПдО.

Приведём для сравнения вариант классического SVD-разложения в тензорной записи. В обозначениях, используемых выше, оно имеет вид

$$A = \sum_{i=1}^{r} \lambda_i u_i v_i^T \quad (5)$$

Псевдообращение: векторный и матричный вариант отображения

7.Определение МПдО через матричное SVD - представление матрицы.

Теорема 3. Для произвольной $m \times n$ - матрицы A её МПдО A_M^+ определяется соотношением

$$A_M^+ X = \sum_{i=1}^{r}\sum_{k=1}^{p} \lambda_i^{-1} v_i e_k^T (u_i e_k^T, X)_{tr} : R^{m\times p} \to R^{n\times p}. \quad (6)$$

8.*Совпадение векторного (классического) ПдО и ПдО матричного*.

Теорема 4. Матричный и векторный вариант ПдО для произвольной $m \times n$ - матрицы A совпадают между собой: т.е. совпадают операторы A^+, A_M^+, определяемые соотношениями (5), (6) соответственно: $A_M^+ = A^+$.

Доказательство. Результат очевидным образом вытекает из теоремы свёртки п.12.

Евклидовы пространства, базовые линейные структуры и ПдО

9. *Основные ортогональные проекторы (ОП), связанные с матрицей линейного оператора*. Основными ОП, связанными с матрицей A являются ОП $P(A^T), P(A)$ на подпространства L_A, L_{A^T} соответственно, а также $Z(A^T), Z(A)$ соответственно - на их ортогональные дополнения в R^m, R^n соответственно. Они определяются соотношениями:

$$P(A^T) = AA^+ = \sum_{i=1}^{r} u_i u_i^T, P(A) = P\left((A^T)^T\right) = A^T(A^T)^+ = A^+A = \sum_{i=1}^{r} v_i v_i^T, \quad (7)$$

$$Z(A) = E_n - P(A) = E_n - A^+A, Z(A^T) = E_m - P(A^T) = E_m - A^{T^+}A^T = E_m - AA^+. \quad (8)$$

Важность последних соотношений определяется тем, что $L_{A^T}^{\perp}$ является множеством нулей оператора A.

Теорема 5. Основными ОП, связанными с матрицей A, как матрицей линейного оператора над матричными евклидовыми пространствами, являются ОП, определяемые соотношениями (7),(8) и являющимися ОП соответственно на $L_A, KerA$.

10. Исследование матричного линейного алгебраического уравнения (МЛаУ).

Теорема 6. Для того чтобы МЛаУ $AX = Y, A \in R^{m\times n} X \in R^{n\times p}, Y \in R^{m\times p}$ было разрешимо необходимо и достаточно, чтобы $trY^T Z(A^T)Y = 0$. В этом случае множество решений Ω_Y определяется соотношением

$$\Omega_Y = \{X : X = A^+Y + Z(A)V, V \in R^{n\times p}\}. \quad (9)$$

В случае, когда МЛаУ несовместно, т.е. когда $trY^T Z(A^T)Y > 0$,множество (9) описывает решение оптимизационной задачи наилучшего квадратического приближения правой части Y значениями левой части AX того же уравнения:

$$\underset{X\in R^{n\times p}}{Arg} \min \| AX - Y \|_{tr}^2 = \Omega_Y = \{X : X = A^+Y + Z(A)V, V \in R^{n\times p}\}. \quad (10)$$

Величина невязки для каждого решения оптимизационной задачи составляет $trY^T Z(A^T)Y$.

Замечание 1.Обратим внимание, что множество решений МЛаУ, и множество псевдорешений, когда МЛаУ несовместно, описываются одним и тем же соотношением (9), (10).

Замечание 2. Поскольку при совместности МЛаУ $\min_{X\in R^{n\times p}} \| AX - Y \|_{tr}^2 = 0$ и этот минимум достигается на решениях МЛаУ, то множество Ω_Y из (9) описывает и множество решений, и множество псевдорешений как решение одной и той же оптимизационной задачи

$$\Omega_Y = \{X : X = A^+Y + Z(A)V, V \in R^{n\times p}\} = \underset{X\in R^{n\times p}}{Arg} \min \| AX - Y \|_{tr}^2 \quad (11)$$

Замечание 3. Обратим внимание, что матрица A^+Y является решением, точным или псевдо -, МЛаУ. Это решение будем называть базовым.

11. Оптимизационное свойство Пенроуза для матричного ПдО.

Теорема 7. Базовое решение МЛаУ как в точном, так и в псевдоварианте, является единственным наименьшим по норме решением оптимизационной задачи поиска наилучшего квадратичного приближения правой части МЛаУ значениями левой части того же уравнения:

$$A^+Y = \underset{X\in \underset{X\in R^{n\times p}}{Arg} \min\|AX-Y\|_{tr}^2}{Arg} \min \| X \|_{tr}. \quad (12)$$

Доказательство. В силу замечания 2 и множество решений, и множество псевдорешений являются решением одной и той же оптимизационной задачи. Поэтому для любого $X \in \underset{x \in R^n}{Arg} \min \| AX - Y \|_{tr}^2$ в соответствии с (10) справедливо представление

$$X = A^+Y + Z(A)V, V \in R^{n \times N}. \tag{13}$$

Поскольку $Z(A)V \in Z(A)R^{n \times p}$ и, как нетрудно убедиться, для любого $V \in R^{n \times p}$

$$A^+Y \perp_{tr} Z(A)V,$$

то в соответствии с многомерным вариантом теоремы Пифагора

$$\| X \|_{tr}^2 = \| A^+Y + Z(A)V \|_{tr}^2 = \| A^+Y \|_{tr}^2 + \| Z(A)V \|_{tr}^2 \geq \| A^+Y \|_{tr}^2.$$

Таким образом, для произвольного $X \in \underset{X \in R^{n \times p}}{Arg} \min \| AX - Y \|_{tr}$

$$\| X \|_{tr}^2 \geq \| A^+Y \|_{tr}^2,$$

и базовое решение A^+Y, псевдо - или точное, является наименьшим по норме решением оптимизационной задачи, определяемой правой частью соотношения (12). Единственность решения оптимизационной задачи вытекает из того, что $\| Z(A)V \|^2 = 0 \Leftrightarrow Z(A)V = 0$.

12. В связи с ограниченностью работы только упомянем другие важные в применениях результаты, которые приведены в работе [Донченко et al, 2010]. Это прямые и обратные формулы Гревиля, а также формулы аналитического возмущения ПдО Н.Ф.Кириченко [Кириченко,1997]. В сущности, в этих результатах речь идёт о формулах, связывающих ПдО изменённой (возмущённой) матрицы с ПдО исходной, а также характеристиками возмущения. В формулах Гревиля таким возмущением является добавление (прямые) или вычёркивание (обратные) строки или столбца матрицы. Собственно, Гревилю принадлежит формула, касающаяся добавления строки, да ещё и только для случая независимости добавляемой строки от строк исходной матрицы. Результаты, связанные с добавлением зависимой строки, как и с вычёркиванием произвольного типа строки, принадлежат Н.Ф.Кириченко [Кириченко,1997].

В аналитических формулах возмущения [Кириченко,1997] изменение исходной матрицы происходит аддитивно: через добавление «простейшей» матрицы. В качестве «простейшей» матрицы выступает тензорное произведение ab^T двух векторов $a \in R^m, b \in R^n$. Как и в случае формул Гревиля, вид ПдО возмущённой матрицы определяется тем, являются ли компоненты возмущения зависимыми или независимыми от, соответственно, столбцов и строк возмущаемой матрицы. Кроме того, результат зависит также и от того, падает или сохраняется ранг возмущённой матрицы, когда составляющие элементы возмущения зависят от соответствующих составляющих возмущаемой матрицы. Обратим внимание, что все формулы или условия описываются явными аналитическими выражениями. Условия линейной зависимости или независимости для столбцов уже нашли своё отражение в п.22. Для описания зависимости от строк в соответствующем условии необходимо только заменить A^T на A.

Что же касается условия падения или сохранения ранга возмущённой матрицы, когда вектор a зависит от столбцов, а вектор b^T - от строк A, то они имеют вид:

$$b^T A^+ a \neq -1 \text{ сохраняется, } b^T A^+ a = -1 \text{ падает.}$$

13. Квадрат расстояния матрицы до гиперплоскости.

Теорема 8. Квадрат расстояния $\rho^2(C, \Gamma(B, L_A))$ матрицы $C \in R^{m \times p}$ от гиперплоскости $\Gamma(B, L_A) = B + L_A$ определяется соотношением

$$\rho^2(C,\Gamma(B,L_A)) = \min_{Y\in\Gamma(B,L_A)} \| C - Y \|_{tr}^2 = tr(C-B)^T Z(A^T)(C-B) .$$

Базовые нелинейные структуры: группирующие операторы

Важнейшими нелинейными структурами евклидова пространства являются квадратичные формы (в работе - неотрицательно определённые) и отвечающие им эллипсоиды или эллипсоидальные цилиндры. Среди таких нелинейных структур принципиальными являются матрицы так называемых «группирующих операторов», которые естественным образом связаны с групповыми свойствами набора векторов. Группирующие операторы возникают в связи с набором векторов $a(j)\in R^m, j=\overline{1,n}$ и отвечающей ему матрицей $A=(a(1)\vdots\ldots\vdots a(n))$. Как и ортогональные проекторы, группирующие операторы являются парными. Будем обозначать их, соответственно $R(A), R(A^T)$. В дальнейшем рассмотрение свойств группирующих операторов будет проводиться в связи с евклидовыми пространствами числовых векторов и переноситься на случай отображений между матричными евклидовыми пространствами

14.*Определение группирующих операторов*. Группирующие операторы, обозначаемые $R(A), R(A^T)$, и для векторного и для матричного варианта действия матрицы A определяются соотношениями

$$R(A^T) = A^{+^T} A^+ , R(A) = A^+ A^{+^T} .$$

Их важность для практики и область применения раскрываются в свойствах, приведённых ниже.

15.*Проектирование на нормированный вектор* $u \in R^m : \| u \| = 1$, *элементов набора векторов* $a(j)\in R^m, j=\overline{1,n}$. Основной результат этого пункта представлен леммой 1 ниже.

Лемма 1. Для произвольного набора $a(j)\in R^m, j=\overline{1,n}$ с матричным представлением $A=(a(1)\vdots\ldots\vdots a(n))$ и произвольного нормированного вектора $u \in R^m : \| u \| = 1$, справедливо равенство:

$$\sum_{j=1}^{n} a^T(j) u u^T a(j) = u^T A A^T u \tag{14}$$

Доказательство. Действительно, принимая во внимание связь п.2 векторов набора $a(j)\in R^m, j=\overline{1,n}$ со своим матричным представлением: $a(j) = Ae_{(j)}, j=\overline{1,n}$, имеем:

$$\sum_{j=1}^{n} a^T(j) u u^T a(j) = \sum_{j=1}^{n} e^T{}_{(j)} A^T u u^T A e_{(j)} = \sum_{j=1}^{n} u^T A e_{(j)} e_{(j)}^T A^T u = u^T A \left[\sum_{j=1}^{n} e_{(j)} e_{(j)}^T \right] A^T u .$$

Остаётся только заметить, что $\sum_{j=1}^{n} e_{(j)} e_{(j)}^T = E_n$, где E_n - единичная матрица в R^n.

Замечание 4. Левая часть соотношения (14) леммы 1 представляет собою сумму квадратов проекций векторов набора $a(j)\in R^m, j=\overline{1,n}$ на нормированный вектор $u \in R^m : \| u \| = 1$.

16.Проектирование на элементы $u_i \in R^m, i=\overline{1,r}$ SVD –разложения матричного представления A набора векторов $a(j)\in R^m, j=\overline{1,n}$. Основной результат этого пункта представлен леммой 2, приведённой ниже.

Лемма 2. Для произвольного набора $a(j)\in R^m, j=\overline{1,n}$, с матричным представлением $A=(a(1)\vdots\ldots\vdots a(n))$ имеет место соотношение:

$$\sum_{j=1}^{n} a^T(j) u_i u_i^T a(j) = \lambda_i^2 , i=\overline{1,r} .$$

Доказательство вытекает из леммы 1 предыдущего пункта и из п.8, в котором наборы $u_i \in R^m, v_i \in R^n, i=\overline{1,r}, r = rankA$ определяются как ортонормированные наборы собственных векторов матриц AA^T, A^TA, отвечающих общему набору ненулевых собственных чисел $\lambda_i^2 > 0, i=\overline{1,r}$.

17. *Группирующие операторы: эллипсоиды группировки набора векторов* $a(j) \in R^m, j=\overline{1,n}$. Основное утверждение пункта – теорема 6 ниже.

Теорема 9. Пусть $a(j) \in R^m, i=\overline{1,n}$ произвольный набор векторов из R^n с матричным представлением $A=(a(1) \vdots \ldots \vdots a(n)), rankA = r \le \min(m,n)$. Тогда все векторы набора принадлежат внутренности эллипсоида, точнее: эллипсоидального цилиндра, определяемого уравнением

$$x^T R(A^T) x = r, x \in R^m,$$

где, $R(A^T)$, группирующий оператор: $R(A^T) = A^{+^T} A^{+}$.

Доказательство. Рассмотрим квадраты проекций векторов набора $a(j), j=\overline{1,n}$, на каждый из векторов $u_i, i=\overline{1,r}$, SVD-представления (5) матрицы A. Принимая во внимание ортонормированность набора $u_i, i=\overline{1,r}$, и обозначая квадраты проекций через $\left\|\Pr_{u_i} a(j)\right\|^2, i=\overline{1,r},\ j=\overline{1,n}$, очевидным образом имеем

$$\left\|\Pr_{u_i} a(j)\right\|^2 = a^T(j) u_i u_i^T a(j), j=\overline{1,n}, i=\overline{1,r}.$$

Суммирование по всем векторам набора $a(j), j=\overline{1,n}$, и применение леммы 2 даёт

$$\sum_{j=1}^{n} a^T(j) u_i u_i^T a(j) = u_i^T AA^T u_i = \lambda_i^2, i=\overline{1,r}.$$

Таким образом, после деления обеих частей последнего соотношения на, соответственно, $\lambda_i^2, i=\overline{1,r}$, имеем

$$\sum_{j=1}^{n} \frac{a^T(j) u_i u_i^T a(j)}{\lambda_i^2} = \frac{u_i^T AA^T u_i}{\lambda_i^2} = 1, i=\overline{1,r},$$

т.е.

$$\sum_{j=1}^{n} \frac{a^T(j) u_i u_i^T a(j)}{\lambda_i^2} = 1, i=\overline{1,r}.$$

Свернув (просуммировав) последнее равенство по $i=\overline{1,r}$, получаем

$$\sum_{i=1}^{m}\sum_{j=1}^{n} \frac{a^T(j) u_i u_i^T a(j)}{\lambda_i^2} = \sum_{i=1}^{m} \frac{u_i^T AA^T u_i}{\lambda_i^2} = r.$$

Поменяв порядок суммирования в двойной сумме, получаем

$$\sum_{j=1}^{n}\sum_{i=1}^{m} \frac{a^T(j) u_i u_i^T a(j)}{\lambda_i^2} = \sum_{j=1}^{n} a^T(j) \sum_{i=1}^{m} \frac{u_i u_i^T}{\lambda_i^2} a(j) = r.$$

Далее, приняв во внимание, что

$$\sum_{i=1}^{r} \frac{u_i u_i^T}{\lambda_i^2} = A^{+^T} A^{+} = R(A^T),$$

получаем окончательно

$$\sum_{j=1}^{n}\frac{a^T(j)u_iu_i^Ta(j)}{\lambda_i^2}=\sum_{j=1}^{n}a^T(j)\sum_{i=1}^{m}\frac{u_iu_i^T}{\lambda_i^2}a(j)=\sum_{j=1}^{n}a^T(j)A^{+^T}A^+a(j)=\sum_{j=1}^{n}a^T(j)R(A^T)a(j)=r\ ,$$

т.е.

$$\sum_{j=1}^{n}a^T(j)R(A^T)a(j)=r\ . \tag{15}$$

Поскольку $R(A^T)$ – симметричная, неотрицательно определённая матрица, то следствием соотношения (15) является одновременное выполнение неравенств

$$a^T(j)R(A^T)a(j)\le r,\ j=\overline{1,n}\ . \tag{16}$$

Та же симметричность и неотрицательная определённость позволяет сделать вывод, что уравнение

$$x^TR(A^T)x=r, x\in R^m \tag{17}$$

определяет эллипсоид, точнее: эллипсоидальный цилиндр, в R^m с длинами $\frac{1}{\lambda_i\sqrt{r}}, i=\overline{1,r}$ нетривиальных полуосей. Напомним, что $r=rankA\le\min(m,n)$.

Таким образом, выполнение неравенства (16) для всех векторов набора $a(j)\in R^m, j=\overline{1,n}$, означает их одновременную принадлежность внутренности эллипсоидального цилиндра с уравнением (17), и доказательство теоремы завершено.

Замечание 5. В действительности неравенство (16) может давать существенное загрубление «радиуса» эллипсоида. Так, при векторах $a(j)\in R^m, j=\overline{1,n}$, близких к ортогональным, очевидным образом, константу в правой части (16) можно выбрать близкой к 1.

18. *Группировка матричных объектов*.

Теорема 10. Пусть $A(j)\in R^{m\times p}, i=\overline{1,n}$ произвольный набор общим рангом r матриц из $R^{m\times p}$, и $\Re$ строчный - кортеж из указанных элементов: $\Re=(A(1),...,A(n))$ Тогда все векторы набора принадлежат внутренности эллипсоида, точнее: эллипсоидального цилиндра, определяемого уравнением

$$trX^TR(\Re^T)X=r, X\in R^{m\times p}\ ,$$

где, $R(A^T)$, группирующий оператор: $R(\Re^T)=\Re^{+T}\Re^+$.

19. *Усиление результата об эллипсоидах группировки для числовых векторов.*

Теорема 11. Все векторы набора $a(j)\in R^m, i=\overline{1,n}$, с матричным представлением $A=\left(a(1)\vdots\ \dots\ \vdots\ a(n)\right)$ принадлежат внутренности эллипсоидального цилиндра

$$x^TR(A^T)x=r_{\max}^2, r_{\max}^2\le r=rankA, x\in R^m\ , r_{\max}^2=\max_{j=\overline{1,n}}a^T(j)R(A^T)a(j)\ ,$$

который будем называть минимальным эллипсоидом группировки для рассматриваемого набора.

Группирующие операторы и расстояние Махаланобиса

Группирующие операторы с точностью до скалярного множителя неожиданным образом оказываются связанными с расстоянием Махаланобиса [Mahalanobis,1936] (см. также, например, [McLachlan, Geoffry, 1992]), и проясняют его суть. Напомним, что расстояние Махаланобиса (в дальнейшем РаМ) является одним из способов определения степени принадлежности $\rho^2(x,N)$ неслучайного вектора $x \in R^m$ многомерному $N(a,B)$ нормальному распределению, и определяется соотношением

$$\rho^2(x,N) = (x-a)^T B^{-1}(x-a). \tag{18}$$

Напомним также, что $a \in R^m$ является математическим ожиданием, а симметричная неотрицательно определённая матрица $B: B \in R^{m\times m}, B^T = B, B \geq 0$, является матрицей ковариаций распределения. Неслучайный вектор $x \in R^m$ интерпретируется как возможный элемент выборки.

Как правило, при использовании РаМ предполагается, что наблюдаемое значение может относиться к одному из нескольких $N_k, k=\overline{1,K}$ нормальных распределений $N(a_k,B_k), k=\overline{1,K}$. В этом случае для каждого из возможных распределений определяется своё РаМ. В этом случае рассматривается набор расстояний вида (18):

$$\rho^2(x,N_k) = (x-a_k)^T B_k^{-1}(x-a_k)\,, \tag{19}$$

а отнесение вектора $x \in R^m$ к одному из K возможных нормальных распределений осуществляется по миниммуму РаМ'ов из (19).

Отметим для сравнения, что средний квадрат расстояния от неслучайного вектора $x \in R^m$ до $N(a,B)$ распределённой случайной величины ξ определяется выражением:

$$\mathrm{M}\|\xi - x\|^2 = trB + \|x-a\|^2,$$

а квадратичная форма $(x-a)^T B(x-a)$ описывает дисперсию случайной величины $(x-a)^T\xi$.

Таким образом, квадратичная форма (18) не является адекватной с точки зрения описания естественных квадратических характеристик нормально распределённого случайного вектора на основе среднего. Единственным обоснованием использования расстояния Махаланобиса яаляется присутствие квадратичной формы из (18) в плотности распределения $f(x), x \in R^m$, определяемой соотношением

$$f(x) = (2\pi)^{-\frac{m}{2}} |B|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(x-a)^T B^{-1}(x-a)\}, x \in R^m.$$

Заканчивая знакомство с РаМ, отметим, что при его практическом использовании вместо параметров a,B используются их оценки $\hat{a},\hat{B}$ на основе выборки $a(1),...,a(n) \in R^m$, определяемые соотношениями

$$\hat{a} = \bar{a} = \frac{1}{n}\sum_{i=1}^{n} a(i)\,, \hat{B} = \frac{1}{n}\sum_{i=1}^{n}(a(i)-\bar{a})^T(a(i)-\bar{a})\,. \tag{20}$$

В этом случае РаМ приобретает «статистическое» представление

$$\hat{\rho}^2(x,N) = (x-\hat{a})^T \hat{B}^{-1}(x-\hat{a}). \tag{21}$$

Теорема 12. Статистический вид РаМ может быть представлен в виде

$$\hat{\rho}^2(x,N) = n(x-\bar{a})^T R(\tilde{A}^T)(x-\bar{a})\,,$$

где $\tilde{A} = (a(1)-\bar{a} \vdots ... \vdots a(n)-\bar{a})$.

Доказательство. Действительно, из (20) непосредственно следует, что $\hat{a} = \bar{a}$. Из тех же соотношений с использованием результата п.3 вытекает, что

$$\hat{B} = \frac{1}{n}\sum_{i=1}^{n}(a(i)-\bar{a})^T(a(i)-\bar{a}) = \frac{1}{n}\tilde{A}\tilde{A}^T \tag{22}$$

Если матрица $\hat{B}$ невырожденная и определяется соотношением (22), то $\hat{B}^{-1} = n(\tilde{A}\tilde{A}^T)^{-1}$. Из соотношения (5) для векторного SVD-разложения для матрицы $\tilde{A}$ вытекает, что

$$\tilde{A}\tilde{A}^T = \sum_{i=1}^{n} \lambda_i^2 u_i^T u_i \ . \tag{23}$$

Из (23) в свою очередь следует, что матрица $(AA^T)^{-1}$, если она существует, определяется соотношением

$$(\tilde{A}\tilde{A}^T)^{-1} = \sum_{i=1}^{m} \frac{u_i^T u_i}{\lambda_i^2} . \tag{24}$$

Как нетрудно убедиться, правая часть в (24) совпадает с $R(\tilde{A}^T)$. Таким образом, окончательно имеем:

$$\hat{B}^{-1} = nR(\tilde{A}^T) \ ,$$

что и завершает доказательство теоремы.

Очевидна, таким образом, эвристическая основа расстояния Махаланобиса и его жёсткая теоретико-вероятностная привязка. Последнее означает, что вне теории вероятностей и математической статистики и вне связи с многомерным нормальным распределением говорить о расстоянии Махаланобиса не имеет смысла.

В то же время группирующие операторы предоставляют возможности выявления групповых свойств векторов на основе минимальных эллипсоид группировки в любой ситуации.

Применения: кластеризация

Примеры применения базовых линейных структур в линейной регрессии, в линейных системах управления с дискретным временем, а также для специального класса функциональных сетей, обобщающих искусственные нейронные сети, можно найти в работах [Кириченко, Донченко,2005] , [Донченко et al 2010].

Отметим дополнительные возможности, появляющиеся в связи с использованием базовых линейных и нелинейных структур для отображений между матричными евклидовыми пространствами. Отметим, что, как и в случае евклидовых пространств числовых векторов, для представлений классов матричных объектов могут использоваться как гиперплоскости, так и эллипсоиды группировки. Заметим. Что использование матричных объектов как «векторов признаков» существенно расширяет границы применимости алгоритмов кластеризации на основе гиперплоскостей или эллипсоидов группировки. Действительно, использование матричных объектов позволяет непосредственно учесть динамику развития объекта через объединение в одну матрицу классических векторов признаков для некоторого дискретизированного интервала времени. Классическим примером представления исследуемого объекта «матричным вектором признаков» является спектрограмма. Напомним, что спектрограмма является матрицей (или её изображением) последовательных наборов энергетических спектров частей звукового сигнала в окне фиксированной длины. Набор энергетических спектров соответствует последовательности окон, которые получаются последовательным и равномерным сдвигом от начала звукового сигнала к его концу. Заметим, что классическое цифровое изображение также является матрицей.

Заключение

В работе предложена и обоснована концепция базовых структур евклидового пространства, как линейных, так и нелинейных. Изложены конструктивные способы описания и взаимного перехода от одних типов структур к другим. В числе других рассмотрены конструктивные способы порождения, описания и использования базовых структур. Упомянутая конструктивность обеспечивается применением как классических результатов ПдО, так и новыми результатами в этой области. В частности, в работе приведена теорема о сингулярном разложении матрицы как линейного оператора между матричными пространствами, доказана теорема свёртки, обеспечивающая эквивалентность векторного и матричного сингулярных и ПдО-матриц на основе соответствующих сингулярных разложений. Рассмотрены применения полученных результатов для построения кластеризации с разными вариантами расстояний соответствия. Введены в рассмотрение кортежные объекты и операции над ними, обобщающие понятия классических числовых векторов и блочных матриц ленточного характера.

Литература

[Донченко, 2010] Донченко В.С. Евклидовы пространства: конструктивные методы описания базовых структур и их использование / Information Models of Knowledge.- Editors: Krassimir Markov, Vitalii Velichko, Oleksy Voloshin - ITHEA.- Kiev, Ukraine – Sofia, Bulgaria. Number 19.– 2010.- P. 362-376.ISBN 978-954-16-0048-1.

[Донченко et al, 2010] Донченко В.,Кривонос Ю., Омардибирова В. Базовые структуры евклидовых пространств: конструктвные методы описания и использования/ New Trends in Classification and Data Mining. – ITHEA, Sofia, Bulgaria. -2010.- ISBN 978-954-16-0042-9.- P. 155-170.

[Донченко. 2009] Донченко В.С. Неопределённость и математические структуры в прикладных исследованиях/ Human aspects of Artificial Intelligence International Book Series Information science & Computing.– Number 12. – Supplement to International Journal “Information technologies and Knowledge”. – Volume 3.–2009. – P. 9-18.

[Донченко, Омардибирова 2005] Донченко В.С., Омардибирова В.Н. Технология классификации электронных документов с использованием теории возмущения псевдообратных матриц// Proceedings of the XI-th International Conference “Knowledge-Dialogue-Solution”. – June 20-30, Varna, 2005. – Volume 1. – C.223-226.

[Кириченко, Донченко, 2008] В.С Кириченко Н.Ф. Донченко В.С. Гиперплоскости в «множествах и расстояниях соответствия»: кластеризация / Artificial Intelligence and Decision Making.– International book series “INFORMATION SCIENCE&COMPUTING”, Number 7.– Sofia 2008.– P. 25-36.

[Moore, 1920] Moore E.H. On the reciprocal of the general algebraic matrix // Bulletin of the American Mathematical Society. – 26, 1920. – P.394 -395.

[Penrose, 1955] Penrose R. A generalized inverse for matrices // Proceedings of the Cambridge Philosophical Society 51, 1955. – P.406-413.

[Алберт, 1977] Алберт А. Регрессия, псевдоинверсия, рекуррентное оценивание. – М.: Наука – 1977. 305 с.

[Кириченко, 1997] Кириченко Н.Ф. Аналитическое представление псевдообратных матриц //Киб. и СА.- №2. –1997.– С.98-122.

[Кириченко, Донченко,2005] Кириченко М.Ф., Донченко В.С. Задача термінального спостереження динамічної системи: множинність розв’язків та оптимізація//Журнал обчислювальної та прикладної математики. – 2005. –№5– С.63-78.

[Кириченко, Донченко, 2007] Кириченко Н.Ф., Донченко. В.С. Псевдообращение в задачах кластеризации// Киб. и СА.- №4, 2007– С.98-122.

[Кириченко, Донченко, 2008] Кириченко Н.Ф. Донченко В.С. Гиперплоскости в «множествах и расстояниях соответствия»: кластеризация / Artificial Intelligence and Decision Making.– International book series “INFORMATION SCIENCE&COMPUTING”, Number 7.– Sofia 2008. – P. 25-36.

[Mahalanobis,1936] Mahalanobis, P C. On the generalized distance in statistics.//Proceedings of the National Institute of Sciences of India.- 1936.-2 (1).-P. 49–55.

[McLachlan, Geoffry, 1992] McLachlan, Geoffry J. Discriminant Analysis and Statistical Pattern Recognition. Wiley Nescience.-1992.- ISBN 0471691151.

Информация об авторе

Владимир С. Донченко *– профессор; Киевский национальный университет имени Тараса Шевченко, факультет кибернетики, Украина, e-mail: voldon@unicyb.kiev.ua*

ЭКСПЕРТНЫЕ МОДЕЛИ МНОГОКРИТЕРИАЛЬНОЙ ОПТИМИЗАЦИИ

Альберт Воронин

Аннотация. *Предложен подход к решению задачи векторной оптимизации сложных технических и экономических систем в тех случаях, когда недостаточны (или отсутствуют) сведения об экспериментально-статистических данных, необходимых для построения регрессионных моделей. Для решения рассматриваемой проблемы предпринимается подход многокритериальной оптимизации с применением нелинейной схемы компромиссов. Приведен модельный пример.*

Ключевые слова*: векторная оптимизация, регрессионные модели, метод экспертных оценок, аппроксимационные полиномы, метод наименьших квадратов, нелинейная схема компромиссов.*

ACM Classification Keywords: H.1 Models and Principles – H.1.1 – Systems and Information Theory; H.4.2 – Types of Systems.

Содержание проблемы

При оптимизации сложных технических и экономических систем часто приходится сталкиваться с тем, что для построения необходимых математических моделей не хватает экспериментально-статистических данных. Положение усугубляется в том случае, когда оптимизация осуществляется по нескольким противоречивым критериям качества.

В условиях острой нехватки экспериментальных данных мы предлагаем получать необходимую информацию («квазиэкспериментальные» данные) от экспертов — специалистов, имеющих достаточный опыт в проектировании и эксплуатации сложных систем рассматриваемого класса.

Исследование проводится на примере векторной оптимизации объектов космической деятельности по обобщенному критерию “надежность-стоимость”, но результаты легко могут быть использованы и в других предметных областях. Под термином "надежность" будем понимать вероятность нахождения определяющих параметров всех элементов объекта в допустимых по условиям работоспособности пределах.

Необходимо учитывать, что в данном случае речь идет о проектировании принципиально новой техники, для которой технико-экономические показатели существенно отличаются от тех, с какими имели дело раньше или работают сейчас разработчики. Одним из специфических аспектов является крайняя ограниченность (а иногда и полное отсутствие) экспериментально-статистических данных, по которым можно было бы определять математические модели надежности и стоимости.

В этих трудноформализуемых условиях приходится прибегать к нетрадиционным подходам, один из которых рассматривается в настоящей работе. Естественно, что в данном случае речь может идти лишь о прикидочных расчетах, об ориентировочном определении основных тенденций при выборе факторов, влияющих на надежность и стоимость разрабатываемых объектов космической деятельности.

Постановка задачи

Для решения задачи оптимизации необходимо иметь следующие отправные данные.

1. Математические модели:

$$y_1^{'} = f_1(x);$$
$$y_2 = f_2(x),$$

где $y_1^{'}$ – надежность объекта космической деятельности (критерий, подлежащий максимизации); y_2 – стоимость мероприятий, от которых зависит надежность (критерий, подлежащий минимизации); f_1 и f_2

– некоторые критериальные функции; $x=\{x_i\}_{i=1}^n$ – *n*-мерный вектор независимых переменных (аргументы оптимизации).

2. Ограничения по независимым переменным $x \in X$, где

$$X=\{x \big| x_{i\min} \le x_i \le x_{i\max,}\ i \in [1,n]\}.$$

3. Ограничения по критериям $y \in M$, где $y=\{y_k\}_{k=1}^s$ – *s*-мерный вектор минимизируемых неотрицательных критериев (в нашем случае *s*=2). Поясним, что в качестве единого способа экстремизации критериев в нашей задаче выбрана минимизация. Чтобы критерий по характеристике надежности сделать тоже минимизируемым, определим

$$y_1 = 1 - y_1^{'}$$

(если стопроцентная надежность выражается единицей). Тогда

$$M=\{y \big| 0 \le y_k \le A_k, k \in [1;2]\}.$$

Ограничения $x_{i\min}, x_{i\max}$ по аргументам $x \in X$ и A_k по критериям $y \in M$ задаются исходя из физических соображений.

Если всё это есть, то имеются все предпосылки для оптимизации космических объектов по критериям надежности и стоимости, т.е. для определения компромиссно-оптимальных значений параметров $x^* = \{x^*_i\}_{i=1}^n$.

Метод решения

Ввиду очевидной противоречивости критериев, необходимо прибегнуть к специфическим методам теории многокритериальной (векторной) оптимизации. Если используется способ скалярной свертки, то математически модель решения задачи векторной оптимизации представляется в виде

$$x^* = \arg\min_{x \in X} Y[y(x)], \tag{1}$$

где $Y(y)$ – скалярная функция, имеющая смысл скалярной свертки вектора частных критериев, вид которой зависит от выбранной схемы компромиссов. При этом нужно убедиться, что ее минимизация приводит к парето-оптимальному решению: $x^* \in X^K$. В работах [1,2] предложена скалярная свертка по нелинейной схеме компромиссов

$$Y[y(x)] = \sum_{k=1}^{s} A_k [A_k - y_k(x)]^{-1}, \tag{2}$$

где *s* – размерность вектора критериев. Свертка (2) дает возможность формализовано получать парето-оптимальные решения, адекватные заданным ситуациям. При *s*=2 модель (1) имеет вид

$$x^* = \arg\min_{x \in X} [\frac{A_1}{A_1 - y_1(x)} + \frac{A_2}{A_2 - y_2(x)}]. \tag{3}$$

Качественный состав вектора х достаточно разнообразен и, соответственно, размерность *n* этого вектора в общем случае велика. Полный учет параметров х привел бы к неоправданному усложнению критериальных функций f_1 и f_2 и к чрезмерным трудностям решения оптимизационной задачи. Поэтому естественным является выбор только наиболее информативных параметров х – координат пространства, в котором будет осуществляться оптимизация критериев y_1 и y_2, в то время как остальные параметры считаются фиксированными и заданными.

Выбор будем выполнять с привлечением экспертов. Их знакомят с условиями задачи, т.е. называют конкретный тип разрабатываемого космического объекта (ракета-носитель или космический аппарат), описывают условия его проектирования, производства, испытаний и эксплуатации. Экспертов просят записать те мероприятия, которые, по их мнению, могут влиять на надежность и стоимость данного космического объекта на разных стадиях жизненного цикла изделия.

Это, например, кратность резервирования систем управления x_1, значения коэффициентов запаса по прочности конструкции x_2 и по мощности энергоисточников x_3; относительный объем входного контроля материалов и комплектующих x_4, подбор технологий производства и относительный объем контроля их стабильности x_5, объем проведения контрольно-выборочных испытаний x_6; объем экспериментальной отработки элементов и систем во всех режимах x_7, величина материального стимулирования персонала x_8 и пр. В результате специальной процедуры [1] определяется адекватный качественный состав и размерность n вектора независимых переменных $x_1, x_2, ..., x_n$ критериальных функций $f_1(x)$ и $f_2(x)$.

Вид критериальных функций зависит от того, какими сведениями располагает исследователь для построения модели. Спектр широк – от полного знания механизмов явлений (детерминированная модель) до полной неопределенности ("черный ящик"). Между этими информационными полюсами находится вероятностный уровень неопределенности. Детерминированную математическую модель $f(x)$ любой характеристики объекта космической деятельности разработать крайне затруднительно ввиду сложности происходящих физических процессов и реакций объекта на комплекс внутренних и внешних факторов.

Рассмотрим, например, критериальную функцию надежности $f_1(x)$ и аппроксимируем ее на множестве аргументов $x \in X$ некоторой приближающей функцией $F_1(x,a)$, известной с точностью до вектора констант (коэффициентов) $a = \{a_j\}_{j=1}^m$.

При выборе вида функции $F_1(x,a)$ нужно иметь в виду следующее. Установлено [1], что наилучшие результаты получаются, если регрессионная модель строится на основе некоторой известной информации о механизмах исследуемых явлений. Тогда модель называется содержательной. Если же такой информации нет, то приходится работать в классе формальных регрессионных моделей и расплачиваться за отсутствие информации большим объемом вычислений.

К формальной модели предъявляются два противоречивых требования. С одной стороны, приближающая функция должна быть достаточно простой, чтобы процессы вычислений не оказались чрезмерно громоздкими. С другой стороны, аппроксимирующая зависимость должна обладать достаточными прогностическими и точностными свойствами. В большинстве практических случаев оба эти требования выполняются в классе регрессионных полиномов второго порядка:

$$F_1(x,a) = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{i,j=1,i<j}^{n} a_{ij} x_i x_j, \tag{4}$$

где a_0, a_1, a_{ij} – коэффициенты. Функция (4) достаточно хорошо адаптируется к топографии целевой функции $f_1(x)$, она способна выражать такие особенности, как овражность и пр. На практике используются различные усечения регрессионного полинома (4), главным образом, линейные приближения.

Определение коэффициентов a может быть выполнено как методами интерполяции, так и по методу наименьших квадратов (МНК). Интерполяционные формулы предусматривают точное совпадение приближающей и целевой функций в опорных точках (узлах интерполяции), количество которых N, как и количество неизвестных констант a, равно m. Коэффициенты a определяются решением определенной системы уравнений для критерия надежности

$$F_1(x^{(u)}, a) = f_1(x^{(u)}), u \in [1, N = m], \tag{5}$$

где $x^{(u)}$ – узлы интерполяции.

Предполагается, что значения целевой функции в узлах аппроксимации $f_1(x^{(u)}), u \in [1, N]$ известны. Определение этих значений с помощью экспертов является ключевым моментом в настоящей работе и рассматривается ниже.

МНК предусматривает N опорных точек (узлов аппроксимации), причем число N может быть больше, меньше или равно (как частный случай) количеству констант m. Неизвестные коэффициенты приближающей функции определяются из условия

$$E(a)=\sum_{u=1}^{N}[F_1(x^{(u)},a)-f_1(x^{(u)})]^2=\min_a. \qquad (6)$$

Используя необходимое условие минимума функции, получим называемую в теории МНК систему *нормальных* уравнений для критерия надежности

$$\frac{\partial E(a)}{\partial a_j}=0,\ j\in[1,m], \qquad (7)$$

решение которой определяет коэффициенты аппроксимирующей функции. Обратим внимание, что независимо от числа выбираемых опорных точек N система нормальных уравнений (7) всегда является определенной.

Для критерия стоимости $f_2(x)$ справедливы все вышеизложенные соображения, но вместо $a=\{a_j\}_{j=1}^m$ в выражении аппроксимирующей функции $F_2(x,b)$ в общем случае фигурирует другой вектор неизвестных констант $b=\{b_h\}_{h=1}^p$.

Спецификой рассматриваемой задачи является то, что получить значения целевых функций в опорных точках очень затруднительно. Действительно, даже для той одной точки, которая соответствует сложившемуся к настоящему времени комплексу мер по обеспечению надежности уже разработанного космического объекта данного класса, нет достаточной статистики для уверенной оценки уровня надежности. Это особенно относится ко вновь разрабатываемым объектам, не имеющим длительного периода эксплуатации. И уж совсем иллюзорны возможности объективной оценки надежности для других точек области существования аргументов оптимизации $x \in X$.

Как всегда в тех случаях, когда задача трудноформализуема, приходится прибегать к методам экспертных оценок. Квалифицированный специалист (эксперт), имеющий достаточный опыт в проектировании, производстве и эксплуатации объектов данного класса, может произвести **мысленный эксперимент** и представить себе, какими будут уровни надежности объекта при различных сочетаниях факторов $x \in X$. Таким образом, в основе метода лежит индивидуальное мнение (постулат), высказываемое специалистом-экспертом об оцениваемой величине, исходя из своего профессионального опыта. Основным недостатком постулирования является субъективность и возможность произвола.

Процедура метода обработки экспертных оценок позволяет уменьшить этот недостаток. Метод заключается в том, что для оценки некоторой количественной характеристики используются постулаты не одного, а нескольких лиц, компетентных в данном вопросе. Предполагается, что "истинное" значение неизвестной нам количественной характеристики находится внутри диапазона оценок экспертов и "обобщенное" коллективное мнение является более достоверным. В работе [1] предложена процедура обработки данных экспертных оценок, в ходе которой получаются уточненные агрегированные оценки, а также (как сопутствующий продукт) коэффициенты доверия к мнению отдельных экспертов.

Применив этот метод к обработке экспертных оценок надежности и стоимости в каждой из N узловых точек области определения $x \in X$, получим два вектора оценок (квазиэкспериментальные данные):

$$\{f_1(x^{(u)})\}_{u=1}^N;$$
$$\{f_2(x^{(u)})\}_{u=1}^N,$$

которые служат основанием для определения векторов констант a и b по условию (5), если применяется способ интерполяции, или по условию (6)-(7), если применяется МНК. Так определяются математические регрессионные модели

$$y_1' = f_1(x)\approx F_1(x);$$
$$y_2 = f_2(x)\approx F_2(x),$$

которые участвуют в оптимизационной процедуре (3). Так как информация о целевых функциях в узловых точках области аппроксимации получена не экспериментально, а путем экспертного оценивания, то и модели $F_1(x)$, $F_2(x)$ называются *экспертными* регрессионными моделями.

Обсудим отдельно проблему выбора способа аппроксимации критериальных функций в заданных обстоятельствах. При различных усечениях регрессионного полинома (4) количество неизвестных констант, как правило, превышает то число N узлов аппроксимации, в которых эксперт может достаточно уверенно дать свою оценку величины критериальной функции. Поэтому, используя способ интерполяции, мы получим недоопределенную систему уравнений, в которой число уравнений меньше, чем число неизвестных констант.

Чтобы выйти из этого положения, следует применить метод наименьших квадратов, который в математике рассматривается как способ решения недоопределенных, переопределенных и определенных (как частный случай) систем уравнений. При этом решение может быть как аналитическим (по формулам (6)-(7)), так и численным, если функция невязок $E(a)$ имеет сложное выражение. В последнем случае используется, например, алгоритм Левенберга-Марквардта [3,4].

Иллюстрационный пример

Рассматривается задача оптимизации процесса разработки перспективной ракеты-носителя по обобщенному критерию «надежность-стоимость». С помощью экспертных процедур [1] выбраны и фиксированы значения факторов, влияющих на надежность изделия, кроме следующих трех (здесь и далее все числовые данные условны):

1. Кратность резервирования систем управления ракетой x_1, выбирается из возможного диапазона от одного до десяти: $x_1 \in [1;10]$;
2. Коэффициент запаса по прочности конструкции x_2, находится в диапазоне от одного до пяти: $x_2 \in [1;5]$;
3. Коэффициент материального стимулирования персонала фирмы-разработчика x_3, может выбираться из диапазона от одного до шести: $x_3 \in [1;6]$.

Требуется, используя противоречивые критерии надежности $y_1^{'} = f_1(x)$ и стоимости $y_2 = f_2(x)$, оценить компромиссно-оптимальные величины этих трех факторов: $x^* = \{x_i^*\}_{i=1}^3$.

Критериальную функцию надежности аппроксимируем линейным приближением со свободным членом

$$y_1^{'} = f_1(x) \approx F_1(x) = a_0 + a_1x_1 + a_2x_2 + a_3x_3,$$

число неизвестных констант $m = 4$.

При определении значений целевой функции в узлах аппроксимации $f_1(x^{(u)}), u \in [1, N]$ будем иметь в виду, что с достаточной степенью уверенности эксперты могут оценить эти значения только в трех (N=3) точках области $x \in X$:

$$x_1^{(1)} = x_{1\max} = 10; x_2^{(1)} = x_{2\max} = 5; x_3^{(1)} = x_{3\max} = 6;$$
$$x_1^{(2)} = x_{1\min} = 1; x_2^{(2)} = x_{2\min} = 1; x_3^{(2)} = x_{3\min} = 1;$$
$$x_1^{(3)} = x_{1nom} = 6; x_2^{(3)} = x_{2nom} = 3; x_3^{(3)} = x_{3nom} = 3.$$

Последняя точка соответствует представлениям экспертов о номинальных значениях факторов надежности.

Пусть эксперты дали следующие свои оценки уровней надежности изделия в указанных узлах аппроксимации (единица соответствует 100-процентной надежности):

$$f_1(x^{(1)}) = 0,8;\ f_1(x^{(2)}) = 0,6;\ f_1(x^{(3)}) = 0,7.$$

Отметим, что в данной задаче N<m и метод интерполяции применен быть не может. Метод МНК предусматривает минимизацию по неизвестным константам функции (6) квадратов невязок, которая при заданных условиях имеет вид

$$E(a) = (a_0 + 10a_1 + 5a_2 + 6a_3 - 0,8)^2 + (a_0 + a_1 + a_2 + a_3 - 0,6)^2 + (a_0 + 6a_1 + 3a_2 + 3a_3 - 0,7)^2.$$

Применение необходимого условия минимума функции (7) приводит к следующей определенной системе нормальных уравнений:

$$3a_0 + 17a_1 + 9a_2 + 10a_3 = 2,1;$$
$$17a_0 + 137a_1 + 69a_2 + 79a_3 = 12,8;$$
$$9a_0 + 69a_1 + 35a_2 + 40a_3 = 6,7;$$
$$10a_0 + 79a_1 + 40a_2 + 46a_3 = 7,5.$$

Это система линейных алгебраических уравнений (СЛАУ), для решения которой разработаны стандартные компьютерные программы. Метод последовательного исключения переменных Гаусса положен в основу программы on-line [5]. Формулы Крамера для решения СЛАУ используются в программе [6]. Применив метод Гаусса для решения нашей системы нормальных уравнений, получим значения неизвестных констант:

$$a_0 = 0,46; a_1 = -0,06; a_2 = 0,25; a_3 = -0,06.$$

Минимизируемый критерий по характеристике надежности определяется по формуле

$$y_1(x) = 1 - y_1^{'}(x) \approx 0,54 + 0,06x_1 - 0,25x_2 + 0,06x_3$$

(имеется ввиду, что единица соответствует 100-процентной надежности изделия).

Аналогичный расчет проведем для критерия стоимости. Критериальную функцию стоимости аппроксимируем линейным приближением без свободного члена:

$$y_2(x) \approx F_2(x^{(u)}, b) = b_1x_1 + b_2x_2 + b_3x_3,$$

где b_1, b_2, b_3 – неизвестные константы (p=3). Оценку уровней стоимости эксперты дали в тех же узлах аппроксимации (N=3), что и в случае критерия надежности. Пусть эти оценки таковы:

$$f_2(x^{(1)}) = 1;\ f_2(x^{(2)}) = 0,3;\ f_2(x^{(3)}) = 0,6$$

(нормированное значение стоимости при максимальных факторах равно единице). Поскольку в данном случае p=N, то для определения констант можно использовать метод интерполяции. Уравнение (5) применительно к критерию стоимости преобразуется к виду

$$F_2(x^{(u)}, b) = f_2(x^{(u)}), u \in [1, N = p],$$

и с учетом числовых данных

$$10b_1 + 5b_2 + 6b_3 = 1;$$
$$b_1 + b_2 + b_3 = 0,3;$$
$$6b_1 + 3b_2 + 3b_3 = 0,6.$$

Решив эту СЛАУ методом Крамера [6], получим

$$b_1 = -0,1; b_2 = 0,3; b_3 = 0,1$$

и выражение для критерия стоимости имеет вид

$$y_2(x) \approx -0,1x_1 + 0,3x_2 + 0,1x_3.$$

Мы получили аналитические выражения критериев надежности и стоимости, что позволяет применить формулу (3) для определения компромиссно-оптимальных значений факторов x* с учетом очевидных ограничений $A_1 = A_2 = 1$:

$$x^* = \arg\min_{x \in X}\left(\frac{1}{1 - 0.54 - 0.06x_1 + 0.25x_2 - 0,06x_3} + \frac{1}{1 + 0,1x_1 - 0,3x_2 - 0,1x_3}\right).$$

Осуществив минимизацию скалярной свертки критериев по аргументам оптимизации методом Нелдера-Мида [1,2], получим

$$x_1^* = 9,99; x_2^* = 3,69; x_3^* = 1,01.$$

Для решения широкого спектра многокритериальных задач разработана программа TURBO-OPTIM [1].

Этот результат иллюстрационного примера можно трактовать следующим образом. При оптимизации разрабатываемой перспективной ракеты-носителя особое внимание следует обратить на резервирование систем управления изделием x_1. Коэффициент запаса по прочности конструкции нужно выбрать ориентировочно немного более середины диапазона $x_2 \in [1;5]$. Не следует возлагать особых надежд на возможности материального стимулирования персонала x_3.

Заключение

Таким образом, данное исследование дает возможность выявить основные тенденции при разработке и оптимизации новой космической техники. Предложенная методика позволила практически решить задачу векторной оптимизации процесса обязательного страхования разрабатываемых объектов космической деятельности [7].

Нужно отдавать себе отчет, что экспертные модели менее информативны, чем регрессионные модели, определенные с помощью реальных экспериментов. Однако, во-первых, экспертные оценки все же отражают, пусть и с возможными искажениями, реальную действительность, а, во-вторых, экспертные модели являются только начальными приближениями и по мере накопления статистических данных поддаются усовершенствованию.

Благодарности

Статья частично финансирована из проекта ITHEA XXI Института Информационных теорий и приложений FOI ITHEA и Консорциума FOI Bulgaria (www.ithea.org, www.foibg.com).

Библиография

1. Воронин А.Н., Зиатдинов Ю.К., Козлов А.И. Векторная оптимизация динамических систем. – Киев: Техніка, 1999. – 284 с.
2. Воронин А.Н. Нелинейная схема компромиссов в многокритериальных задачах оценивания и оптимизации // Кибернетика и системный анализ – 2009. – №4.– С. 106-114.
3. Levenberg K. A method for the solution of certain problems in least squares // Quart. Appl. Math., 1944, Vol. 2, pp. 164-168.
4. Marquardt D. An algorithm for least-squares estimation of non-linear parameters // SIAM J. Appl. Math., 1963, Vol. 11, pp. 431-441.
5. http://www.webmath.ru/web/prog13_1.php.
6. http://www.webmath.ru/web/prog12_1.php.
7. Воронин А.Н., Кириченко А.А., Козлов А.И. Регрессионная модель экспертных оценок «надежность-стоимость» в задачах векторной оптимизации страхования объектов космической деятельности // Проблемы управления и информатики – 1999. – №3. – С. 58-65.

Сведения об авторе

***Воронин Альберт Николаевич** – профессор, доктор технических наук, профессор кафедры компьютерных информационных технологий Национального авиационного университета, проспект Комарова, 1, Киев-58, 03058 Украина; e-mail:* alnv@voliacable.com

МНОГООСНОВНЫЕ АЛГЕБРЫ, АБСТРАКТНЫЕ ТИПЫ ДАННЫХ И ТРАНСФИНИТНАЯ РЕКУРСИЯ

Крывый С.Л.

Аннотация: *Описывается абстрактный тип данных ``натуральное число" как многоосновная алгебраическая система, пополненная операторами рекурсии и индукции. Исследуются пределы, до которых распространяется действие рекурсии и индукции на множества, отличные от множества натуральных чисел.*

Ключевые слова: *многоосновные алгебры, абстрактный тип данных, трансфинитная рекурсия.*

Keywords: MSC, verification, RAD.

ACM Classification Keywords: D.2.4 Software/Program Verification - Formal methods

Введение

Абстрактные типы данных (АТД) активно используются в процессе разработки программного обеспечения, его обосновании и верификации. Этот упрощенный процесс можно схематически изобразить в таком виде [axo-79]:

Математическая модель		Абстрактный тип данных		Структуры данных
----------------------------------	$\Rightarrow$	----------------------------------	$\Rightarrow$	---------------------------
Неформальный алгоритм		Программа на псевдокоде		Программа на языке программирования

С помощью таких технологий строятся системы баз данных, баз знаний, системы компьютерной алгебры и геометрии, графические системы и т. д. Главным требованием при построении таких систем является правильность и эффективность построенных алгоритмов и программ. В процессе построения эффективных алгоритмов и программ важную роль играет успешный выбор структур данных. Неформально описывая алгоритм решения прикладной задачи, исходя из выбранной математической модели, приходится пользоваться такого типа данными, которых нет ни в одном языке программирования, но которые свойственны этой математической модели. Такого типа данные называются *абстрактными типами данных*. Более формально, под абстрактным типом данных понимают некоторую формальную математическую модель вместе с операциями, функциями и предикатами, определёнными на этой модели. Примерами такого типа данных могут служить множества, вместе с операциями объединения, пересечения и разности. В модели АТД операторы могут иметь операндами не только данные, которые определяются этим АТД, но и операнды языка программирования и операнды, которые определены другими АТД. Результатом выполнения оператора тоже может быть тип данных, который не определяется данным АТД. Но в рамках данного АТД считается, что хотя бы один операнд или результат любого оператора имеет тип данных, определённый в данной модели АТД.

АТД определяются, как правило, с помощью аксиоматики и отличаются от структур данных в языках программирования тем, что при использовании АТД абстрагируются от способа их реализации и рассматриваются только их свойства, которые вытекают из аксиоматики. В общем случае определение АТД сводится к определению таких понятий:

- объекты,
- базовые операции,

- конструкторы новых операций, функций и предикатов,
- определения (дефиниции) новых объектов.
- Основными конструкторами, которые используются при определении новых операций, функций и предикатов есть рекурсия и индукция. Неформально под *рекурсией* понимают такой способ определения функции, при котором значение функции для произвольных значений ее аргументов выражаются известным способом через значения этой функции для меньших значений ее аргументов. Простейшим видом рекурсии является *примитивная рекурсия*.
- *Индукцией* называют такой способ определения функции, при котором значения функции для произвольного значения ее аргумента выражаются известным образом через значения этой функции для аргумента, которые непосредственно предшествует данному аргументу.
- Известно, что индукция, как и рекурсия работают на специального типа множествах, которые являются полностью упорядоченными. Основной вопрос, который нас будет интересовать - это пределы до которых можно распространить действие рекурсии и индукции.
- Рассмотрим АТД “натуральное число”, операции и предикаты, определённые на нем, а также определение новых операций и предикатов на этих АТД с помощью индукции и рекурсии. Этот АТД описывается в виде многоосновной алгебраической системы. Заметим, что каждую алгебраическую систему [Мальцев-70] можно рассматривать как многоосновную алгебру. Действительно, если

$$A = (\mathrm{A}, \Omega = \left\{\omega_1^{k_1},...,\omega_n^{k_n}\right\}, \Pi = \left\{\pi_1^{m_1},...,\pi_r^{m_r}\right\}), \quad i = 1,2,...,n,\ j = 1,2,...r,$$

- алгебраическая система, где $w_i^{k_i} : A^{k_i} \to A$, $\pi_j^{m_j} : A^{m_j} \to \{true(1),\ false(0)\}$, а *true* и *false* булевские константы (обычно обозначаемые как 1 и 0), то с этой системой ассоциируется двухосновная алгебра

$$A = (U = (A, \{true,\ false\}), \Omega = \{\omega_1^{k_1},...,\omega_n^{k_n}, \pi_1^{m_1},...,\pi_r^{m_r}\}).$$

- В этой многоосновной алгебре каждая операция $w_i^{k_i}$ имеет тип элементов множества *A*, а каждая операция $\pi_j^{m_j}$ имеет тип булевский. Таким образом при рассмотрении АТД как алгебраической системы мы имеем дело с многоосновными алгебрами.

Алгебраическая система «натуральное число»

Аксиоматика. Рассмотрим АТД *N*, элементы которого называются натуральными числами. Как известно этот тип данных определяется аксиоматическим способом:

1. $0 \in N$;

2. $s : N \to N$ - унарная операция, называемая “следующее натуральное” и удовлетворяет условию: если $n \in N$, то $s(n) \in N$;

3. $\forall n \in N\ s(n) \neq 0$;

4. если $s(n) = s(m)$, то $m = n$;

5. если *A* - множество такое, что $0 \in A$ и из того, что для произвольного $n \in A$ вытекает $s(n) \in A$, то $A = N$.

Это известная *аксиоматическая система Пеано* и ее алгебраическая система принимает вид:

$$A = (N, \Omega = \{0, s\}, \Pi = \{=\}),$$

где *N* - носитель алгебраической системы, *0, s* – нульарная и унарная операции на *N* (причём, операция *s(x) = x + 1*), а = - бинарный предикат равенства, определённый на *N*.

Наличие непустых множеств операций и предикатов, определённых на множестве *N*, позволяет ввести в рассмотрение множества термов (алгебраических выражений) и формул, пользуясь традиционными булевыми связками.

Сначала заметим, что каждое натуральное число можно получить с помощью операций *0* и *s*. Действительно, пусть $m \in N$ - любое натуральное число, тогда

$$m = \underbrace{s(...s(s}_{m \text{ раз}}(0)))...),$$

где операция *s* применяется *m* раз к числу *0.*

Аксиома 4 использует кроме операции *s(n)* также операцию *pr(n)*, которая называется "предшественник" числа *n*, причем *pr(0)=0*. Эти операции связаны между собой очевидными соотношениями *pr(s(n)) = n* и *s(pr(n)) = n*, если $n \neq 0$.

Множество термов T_N определяется индуктивно.

Определение 1. Термами называются выражения, построенные по таким правилам:

а) *0* - терм,

б) если *n* терм, то *s(n)* и *pr(n)* - термы,

в) термами есть те и только те выражения, которые построены по правилам а) - б).

Множество формул F_N определяется индуктивно аналогичным образом.

Определение 2. Формулами называются выражения, построенные по таким правилам:

а) *true, false* - формулы,

б) если *m, n* - термы, то *m = n* - формула,

в) если *A* и *B* - формулы, то $A \wedge B, A \vee B, A \to B$ - формулы,

г) формулами являются те и только те выражения, которые построены по правилам а) - в).

Пользуясь этими множествами, вышеприведенную алгебраическую систему можно расширить таким образом:

$$\mathrm{A} = (U = (T_N, F_N), \Omega = \{0, s, pr, \neg, \vee, \wedge, \to\}, \Pi = \{true,\ false, =\}),$$

где $s, pr : T_N \to T_N$, $=: T_N \times T_N \to F_N,$

$\vee, \wedge, \to: F_N \times F_N \to F_N$, $\neg : F_N \to F_N$.

Следовательно, полученная таким способом АС, является двухосновной и ее называют ***моделью Пеано***. Используя индуктивное определение множества натуральных чисел *N*, можно ввести операцию декартового произведения $N \times N.$

Элементами множества $N \times N$ являются те и только те элементы, которые построены по правилам:

а) $(0,0) \in N \times N,$

б) если $(m,n) \in N \times N,$ то *(s(m),n)* и *(m,s(n))* $\in N \times N.$

Это определение очевидным образом обобщается на произвольные множества *A, B,…,C* и функцию *f*, которые имеют индуктивные определения.

Определение традиционных операций и предикатов

Операции. Напомним сначала схему операции примитивной рекурсии, с помощью которой определяются функции. Будем говорить, что *(n+1)*-арная функция *f* получена из *n*-арной функции *g* и *(n+2)*-арной функции *h* с помощью ***операции примитивной рекурсии***, если

$$f(x_1, x_2, ..., x_{n-1}, x_n, 0) = g(x_1, x_2, ..., x_n);$$

$$f(x_1, x_2, ..., x_{n-1}, x_n, s(y)) = h(x_1, x_2, ..., x_n, y, f(x_1, x_2, ..., x_n, y)).$$

Частным случаем примитивной рекурсии является ***итерация***, схема которой имеет вид [Гудст-70]:

$$f(x_1, x_2, ..., x_{n-1}, x_n, 0) = g(x_1, x_2, ..., x_n);$$

$$f(x_1, x_2, ..., x_{n-1}, x_n, s(y)) = h(x_1, x_2, ..., x_n, f(x_1, x_2, ..., x_n, y)).$$

Введем операцию сложения натуральных чисел “*plus*”, пользуясь операциями полученной АС и рекурсии. Это определение имеет вид:

plus(0,y) = y;

plus(s(x),y) = s(plus(x,y)).

Можно ввести операцию сложения и другим способом:

plus(0,y) = y;

plus(s(x),y) = plus(x,s(y))).

С помощью операции *plus*, вводится операция умножения *mult (x,y)* натуральных чисел *x* и *y:*

mult(0,y) = 0;

mult(s(x),y) = plus(mult(x,y),y)).

Используя операции *s* и *pr*, введенные выше операции *plus* и *mult* принимают вид:

plus(x,y) = if x = 0 then y else s(plus(pr(x),y)),

mult(x,y) = if x = 0 then 0 else plus(mult(pr(x),y),y).

Из этих определений вытекает, что в обоих операциях присутствует унарный предикат *x = 0*. Расширим сигнатуру предикатов АС этим предикатом, который обозначим *isz : N* $\rightarrow$ *{true, false}* (is zero). Теперь ***алгебраическая система N*** принимает вид:

$$A = (U = (T_N, F_N), \Omega = \{0, s, pr, \neg, \vee, \wedge, \rightarrow\}, \ \Pi = \{true, \ false, =, isz\}),$$

с *множеством аксиом* для введенных операций *pr(x)* и предиката *isz(x):*

$$\neg(true) = false, \qquad \neg(false) = true,$$

$$isz(0) = true; \qquad isz(s(x)) = false;$$

$$pr(0) = 0; \ \ pr(s(x)) = x; \qquad \neg(isz(x)) \rightarrow (s(pr(x)) = x).$$

В новой системе аксиомы 3 и 4 принимают вид: *isz(s(x)) = false* и *pr(s(x)) = x* соответственно. Действительно, из того что *s(x) = s(y)* вытекает *pr(s(x)) = x = pr(s(y)) = y*. Введем ещё некоторые функции и предикаты, пользуясь возможностями этой АС:

$$sg(m) = \begin{cases} 0, если \ m = 0, \\ 1, если \ m > 0, \end{cases} \quad \overline{sg}(m) = \begin{cases} 1, если \ m = 0, \\ 0, если \ m > 0, \end{cases} \quad monus(m,n) = \begin{cases} m - n, если \ m \geq n, \\ 0, если \ m < n. \end{cases}$$

Хорошо известно, что эти функции можно получить операцией примитивной рекурсии [Мальцев-70]. Заметим также, что операция *pr(m)* - примитивно рекурсивная функция, поскольку

pr(0) = 0, pr(s(m)) = m.

Предикаты. Пользуясь этими функциями, получаем выражения для функции |m - n| и предикатов m = n (eq(m,n)), m $\leq$ n (leg(m,n)), m < n (less(m,n)):

|m - n| = monus(m,n) + monus(n,m), eq(m,n) = $\overline{sg}$ (|m-n|) = $\overline{sg}$ (monus(m,n) + monus(n,m)),

leq(m,n) =monus(1,monus(m,n)), less(m,n) = $\overline{sg}$ ($\overline{sg}$ (monus(n,m))).

Нетрудно показать, что эти функции и предикаты удовлетворяют таким соотношениям:

monus(0,m) = 0, |m - 0| = |0 - m| = m,

eq(0,s(n)) = eq(s(m),0) = false, eq(0,0) = true,

leq(0,s(n)) = true, leq(s(m),0) = false, leq(0,0) = true,

less(0,s(n)) = true, less(s(m),0) = false, less(0,0) = false

и если *m >0, n> 0*, то

monus(m,n) = monus(pr(m),pr(n)),

|m - n| = |pr(m) - pr(n)|,

eq(m,n) = eq(pr(m),pr(n)),

leq(m,n) = leq(pr(m),pr(n)),

less(m,n) = less(pr(m),pr(n)).

Эти соотношения стают полезными при вычислении значений этих функций и предикатов.

Заметим, что в этой АС предикат eq (предикат =) становится производным, т .е. выражается через предикат *isz*. Исходя из этого, АС *N* принимает более лаконичный вид:

$$A = (U = (T_N, F_N), \Omega = \{0, s, pr, \neg, \vee, \wedge, \rightarrow\}, \Pi = \{true, false, isz\}).$$

Операции *plus* и *mult* в этой АС принимают вид:

plus(m,n) = if isz(m) then n else s(plus(pr(m),n)).

mult(m,n) = if isz(m) then 0 else plus(mult(pr(m),n),n).

Из определения *leq* ($\leq$) вытекают такие свойства:

а) (рефлексивность *leq*). Действительно, *leq(m,m) = true*, и в силу свойств этого предиката получаем *leq(m,m) = leq(0,0) = true*.

б) (антисимметричность *leq*). Если *leq(m,n) = true* и *leq(n,m) = true*, то в силу свойств этого предиката получаем: *leq(m,n) = true* и *leq(n,m) = true* тогда и только тогда, когда *leq(0,n') = true* и *leq(n',0) = true*, но это возможно только в том случае, когда *n' = 0*, а это значит что *m = n.*

в) (транзитивность leq). Если leq(m,n) = true и leq(n,k) = true, то leq(m,n) = true тогда и только тогда, когда leq(0,n') = true, где n' $= \underbrace{pr(...pr(pr}_{m \text{ раз}}(n))...)),$ и leq(n,k) = true $\Leftrightarrow$ leq(0,k') = true, где k' $= \underbrace{pr(...pr(pr}_{n \text{ раз}}(k))...)).$ Тогда leq(m,k) = true $\Leftrightarrow$ leq(0,k'') = true, где

k'' $= \underbrace{pr(...pr(pr}_{m \text{ раз}}(k))...)),$ и *leq(n',k'') = true* в силу *leq(n,k) = true*.

г) Из вышеприведённых свойств предиката вытекает, что любые два элемента множества *N* сравнимы. Следовательно, отношение *leq* является отношением линейного порядка. Более того, из аксиомы 5 вытекает, что множество *N* является полностью упорядоченным, т. е. произвольное непустое его подмножество имеет минимальный элемент. Действительно, предположим что существует подмножество *B* множества *N*, в котором нет минимального элемента. Рассмотрим множество *C = N \ B*. Та как $0 \in C$ (в противном случае 0 был бы минимальным в *B*), то *s(0)* $\in$ *C* на том же основании. Но тогда в множестве *C* нет ни одного элемента, поскольку *C = N* в силу аксиомы 5. А отсюда следует, что *B* = $\varnothing$.

С помощью операции *mult* вводится отношение *делимости* натуральных чисел.

Определение 3. Натуральное число m называется кратным натуральному числу x, если существует такое натуральное число y, что m = mult(x,y). Число x в этом случае называется делителем числа m (обозначение m|x).

Далее, с помощью введённых выше предикатов вводятся бинарные функции *floor(m,n)* - наибольшее целое число, не превосходящее частного от деления *m* на *n*), *ceil(m,n)* - наименьшее целое число, не меньшее частного от деления *m* на *n*), а также операция mod*(m,n)* - частное от деления *m* на *n* (*n > 0*):

$$floor(m,n)=\begin{cases}0, if\ \ m=0,\\ \sum_{i=1}^{m}\overline{sg}\ (\ mult\ (i,n),\ m),\ in\ other\ case.\end{cases}$$

$$ceil(m,n)=\begin{cases}0,\ if\ \ m=0,\\ s(\sum_{i=1}^{m}\overline{sg}\ (\ mult\ (i,n),\ m)),\ \ in\ other\ case.\end{cases}$$

$$\mathrm{mod}(m,n)=\begin{cases}m, if\ \ less(m,n)=true,\\ \mathrm{mod}\ (\ monus\ (m,n),\ n),\ in\ other\ case.\end{cases}$$

Из этих определений вытекают такие простейшие свойства введённых функций:

floor(0,n) = 0, ceil(0,n) = 0,

floor(m,n) = *ceil(m,n)* тогда и только тогда, когда *m*|*n*,

ceil(m,n) - floor(m,n) = 1, если m $\neq$ 0,

mod(0,n) = 0, m = mult(n,floor(m,n)) + mod(m,n),

mod(m,n) = m/n – mult(n,floor(m,n)), если n $\neq$ 0.

Свойства операций, предикатов и отношений

Операции, предикаты и отношения, определённые на данной АС, имеют хорошо известные свойства. Прежде всего это предикат равенства, который является рефлексивным, симметричным, транзитивным и удовлетворяет свойству подстановки, т. е. $\forall x, y, z \in N$

(PE) x = x, (CE) x = y $\rightarrow$ y = x, (TE) (x = y $\wedge$ y = z) $\rightarrow$ x = z,

(SE) (t = t' $\wedge$ f(...t...)) $\rightarrow$ f(...t'...).

Операции *plus* и *mult* удовлетворяют законам коммутативности, ассоциативности и дистрибутивности. Т. е. имеет место

Теорема 1. а) plus(x,y) = plus(y,x), б) plus(x,plus(y,z)) = plus(plus(x,y),z),

в) mult(x,y) = mult(y,x), г) mult(x,mult(y,z)) = mult(mult(x,y),z),

д) mult(x,plus(y,z)) = plus(mult(x,y),mult(x,z)).

Доказательство оставляем читателю.

На основе коммутативности операции *mult,* получаем, что когда число *x* является делителем числа *m*, т. е. существует такое число *y*, что *m* = *mult(x,y)*, то и число *y* тоже является делителем числа *m*. В частности, отсюда вытекает, что 1 и *m* являются делителями *m*.

Отношение делимости рефлексивное, транзитивное и антисимметричное:

а) m |m (рефлексивность),

б) m|n $\wedge$ n|k $\rightarrow$ m |k (транзитивность),

в) m|n $\wedge$ n|m $\rightarrow$ m = n (антисимметричность).

В самом деле, в случае а) имеем:

mult(1,m) = plus(mult(pr(1),m),m) = plus(mult(0,m),m) = plus(0,m) = m.

В случае б) на основе теоремы 1.в) и 1.г) получаем:

m = mult(n,x) $\wedge$ n = mult(k,y) $\rightarrow$ mult(mult(k,y),x) = mult(k,mult(y,x)).

Отсюда следует, что $m|k$. В случае в) имеем $m = mult(n,k) \wedge n = mult(m,s)$. Отсюда получаем:

m = mult(mult(m,s),k) = mult(mult(s,m),k) = mult(s,mult(m,k)) = plus(mult(pr(s),mult(m,k)),mult(m,k)).

Это равенство выполняется только в том случае, когда pr(s) = 0 и k = 1, т. е. s = k = 1. Но тогда n = mult(m,1) = mult(1,m) = plus(mult(0,m),m) = plus(0,m) = m.

На основе рефлексивности отношения делимости вводятся понятия простого числа и общего делителя. Число *m* называется *простым*, если оно имеет всего два делителя 1 и *m*. Пусть $m, n \in N$ и существует $x \in N$ такое, что $m|x$ и $n|x$. В этом случае число *x* называется *общим делителем* чисел m и n. Наибольшее среди общих делителей число *k* называется *наибольшим общим делителем* чисел *m* и *n* (*НОД(m,n)*) Из полученных свойств отношения делимости вытекают следующие свойства НОД:

НОД(m,m) = m,

НОД(m,n) = НОД(m-n,n), если m > n и

НОД(m,n) = НОД(m,n-m), если m < n.

Отсюда, как правило, начинаются курсы по теории чисел и формальной арифметике.

Трансфинитная индукция и рекурсия

Рассмотренная алгебраическая система имеет носителем полностью упорядоченное множество (пум) натуральных чисел *N*. Возникает вопрос: на каждое ли полностью упорядоченное множество (пум) распространяется принцип математической индукции и рекурсии?

Полностью упорядоченные множества. Прежде чем давать ответ на этот вопрос, рассмотрим структуру пум. Непосредственно из определения пум вытекает, что пум имеет наименьший элемент. Кроме того,

- для каждого элемента *x*, кроме максимального, в пум существует элемент *y*, который является непосредственно следующим для *x*, т. е. если *y* непосредственно следует за *x*, то не существует элемента *z*, для которого выполняются неравенства $y > z > x$ (отношение доминирования в пум); элемент, который непосредственно следует за элементом *x* обозначается $x + 1$, за ним $x + 2$ и т. д. (аналог операции *s(x)* на множестве *N*);

- некоторые элементы в пум могут не иметь непосредственного предшественника (например, если $N \times N$ есть множеством, элементы которого упорядочены отношением лексикографического порядка $(m,n) \le (m',n') \Leftrightarrow (m < m') \vee ((m = m') \wedge (n \le n'))$, то элементы *(n,0)* не будут иметь предшественника и такие элементы называются *предельными*;

- произвольный элемент пум имеет вид $x + n$, где *x* - предельный элемент, а *n* - натуральное число (обозначение $x + n$ следует понимать в смысле первого из этих пунктов). Действительно, если *x* не предельный элемент, то возьмём его непосредственного предшественника *pr(x),* и если он не предельный, то возьмём его непосредственного предшественника *pr(pr(x))* и т. д. Бесконечно этот процесс длится не может в силу полной упорядоченности множества. Понятно также, что такое представление элементов пум однозначно, поскольку в каждого элемента может быть не более одного предшественника. Имеет место такое очевидное утверждение.

Теорема 2. Произвольное ограниченное сверху пум имеет наибольший элемент.

Доказательство. Множество *A* называется ограниченным сверху, если существует элемент $a \in A$ такой, что для всех $x \in A$ выполняется неравенство $x \le a$. Поскольку любой элемент пум имеет вид $x + n$, то когда этот элемент имеет непосредственно следующего, то выбираем этот следующий и т. д. Этот выбор не может быть бесконечным, поскольку множество *A* ограничено сверху. Последний из выбранных элементов, для которого в множестве *A* не существует следующего за ним элемента и будет искомым элементом. □

Одним из способов построения пум является операция декартового произведения. Рассмотрим декартово произведение $N \times N$. На этом множестве можно определить разные порядки. Чаще всего встречаются покоординатный порядок и лексикографический порядок, о котором говорилось выше. Определение покоординатного порядка имеет вид:

(m,n) $\le$(m',n') $\Leftrightarrow$ (m $\le$ m') $\wedge$ (n $\le$ n'),

а определение лексикографического выглядит так:

(m,n) $\le$ (m',n') $\Leftrightarrow$ (m < m') $\vee$ (m = m') $\wedge$ (n $\le$n').

Первый из этих порядков частичный, а второй линейный и, даже, полный. Он имеет вид *(0,0)* < *(0,1)* < *(0,2)* < … < *(1,0)* < *(1,1)* < … < *(2,0)* > *(2,1)* < … . Как отмечалось, у элементов вида (*n,0*) нет предшественников и они являются предельными. Общий случай построения пум вытекает из такого утверждения.

Теорема 3. Если множество A полностью упорядочено, то декартово произведение A$\times$A тоже полностью упорядоченное множество относительно лексикографического порядка.

Доказательство. Пусть *B* $\subseteq$*A*$\times$*A, (a,b)* $\in$ *B* = $A_1 \times A_2$, где $A_1, A_2 \subseteq$*A* и a_0, b_0 - наименьшие элементы в множествах A_1, A_2 соответственно. Если $a \ne a_0$ и $b \ne b_0$ (иначе *(a,b)* - наименьший элемент в множестве *B*), то рассмотрим последовательность элементов

(a,b), (pr(a),b), (pr(pr(a)),b),…,(a_0*',b), (* a_0*',pr(b)), (* a_0*',pr(pr(b))),…, (* a_0, b_0*).*

Эта последовательность не может быть бесконечной в силу полной упорядоченности множеств *A*. Тогда элемент (a_0, b_0) - наименьший элемент в множестве *B*. □

Примером такого типа пум может служить множество всех слов *F(X)* конечной длины в некотором конечном алфавите *X*, при условии, что символы алфавита *X* линейно упорядочены. Свойства этого множества детально исследовались в работе [Крывый-10].

Следующая теорема даёт ответ на поставленный выше вопрос.

Теорема 4 [Принцип трансфинитной индукции]. Пусть a_0 - наименьший элемент в пум A и P(x) – некоторое свойство элемента x $\in$A. Тогда, если из истинности P(a_0) и P(x) для всех x < a вытекает истинность P(a), то P(x) истинно для всех x из A.

Доказательство. Допустим противное, т. е. что существует такое непустое подмножество *A'* элементов из *A*, что *P(a*) ложно на элементах этого множества при выполнении условий теоремы. Пусть *a* - минимальный элемент в *A'*. Поскольку *P(* a_0*)* истинно, то *a* $\ne$ a_0 и *a* > a_0. Из условий теоремы вытекает, что *P(x)* истинно для всех *x* < *a*, но тогда из этих же условий должна следовать истинность и *P(a),* а это противоречит нашему предположению. □

Перейдём к рассмотрению рекурсии. Сначала введём понятие начального отрезка линейно упорядоченного множества. Если линейно упорядоченное множество *A* разбито на два непересекающиеся подмножества *B* и *C* так, что любой элемент множества *B* меньше произвольного элемента множества *C*, то множество *B* называют *начальным отрезком множества A*.

Пусть *A* - произвольное пум. Поскольку пум является линейно упорядоченным множеством, которое имеет минимальный элемент, то будем обозначать этот элемент 0, а произвольный его начальный отрезок - *[0,x)* или *[0,x].* Мы хотим дать рекурсивное определение функции *f : A* $\to$*B*, где *B* - некоторое множество. Такое определение должно связывать значения функции *f(x)* на некотором элементе *x* $\in$*A* со значениями *f(y)* для всех *y* < *x*. Это значит, что рекурсивное определение определяет функцию *f(x)* в предположении, что известно ограничение функции *f* на начальный отрезок *[0,x)* множества *A*. Обозначим $f\,|_{[0,x)}$ сужение функции *f* на начальный отрезок *[0,x).*

Теорема 5. Пусть A - пум, B - произвольное множество и задано некоторое рекурсивное правило, т.е. функция F, которая ставит в соответствие элементу $x \in A$ и функции g: $[0,x) \to B$ некоторый элемент множества B. Тогда существует единственная функция $f: A \to B$, для которой $f(x)=F(x, f|_{[0,x)})$ для всех $x \in A$.

Доказательство теоремы вытекает из такого утверждения: существует единственное отображение *f* отрезка *[0,a]* в множество *B*, для которого рекурсивное определение $f(x) = F(x, f|_{[0,x)})$ выполняется для всех $x \in [0,a]$. Такое отображение будем называть *корректным*. Следовательно, необходимо доказать, что для каждого $a \in A$ существует единственное корректное отображение *f* отрезка *[0,a]* в множество *B*.

Допустим, что для всех *c<a* утверждение имеет место. Это значит, что существует единственное корректное отображение $f_c: [0,c] \to B$, т. е. что для всех $d \le c$ значения f_c совпадают со значениями, которые получены на основе рекурсивного определения.

Рассмотрим отображения f_{c_1} и f_{c_2} для двух разных значений c_1 и c_2. Пусть, например, $c_1 < < c_2$. Отображение f_{c_2} определено на большем отрезке $[0, c_2]$ и если f_{c_2} ограничить на меньший отрезок, то оно совпадет с f_{c_1}, поскольку ограничение корректного отображения на меньший отрезок будет корректным отображением в силу предположения о единственности отображения на отрезке $[0, c_1]$.

Таким образом, все отображения f_c согласуются между собой в том смысле, что когда их значения определены, то они одинаковы. Объединяя их, получаем некоторое единственное отображение *h*, которое определено на отрезке *[0,a)*. Применяя к *a* и *h* рекурсивное правило, получаем некоторое значение $b \in B$. Доопределим *h* в точке *a*, полагая *h(a) = b*. В результате получаем отображение $h : [0,a] \to B$, которое, очевидно, является корректным.

Покажем, что на отрезке *[0,a]* корректное отображение единственно. Действительно, его ограничение на отрезок *[0,c]* при *c < a* должно совпадать с f_c и остаётся показать, что в точке *a* это отображение однозначно. Но последнее гарантируется рекурсивным определением. Этим индуктивное доказательство завершается.

Остаётся заметить, что для разных *a* корректные отображения согласуются между собой и поэтому задают некоторую функцию $f : A \to B$, которая удовлетворяет рекурсивному определению.

Таким образом, доказано существование и единственность функции, поскольку ограничение этой функции на произвольный отрезок *[0,a]* корректно и поэтому определено однозначно. □

Аналогичная ситуация возможна и в общем случае, т. е. в случае когда рекурсивное правило не полностью определено. Введём общее понятие согласованности функций (частный случай которого фигурировал в доказательстве предыдущей теоремы), заданных на одних и тех множествах.

Определение 4. Пусть *A* - пум, а *B* - произвольное множество. Частичные функции $f : A \to B$ и $g : A \to B$ называются *согласованными функциями*, если *f(x)* и *g(x)* или одновременно не определены, или определены и *f(x) = g(x),* где $x \in A$.

Имеет место

Теорема 6 [Принцип трансфинитной рекурсии]. Пусть $f(x) = F(x, f|_{[0,x)})$ является частичным отображением, т. е. для некоторых x и g : $[0,x) \to B$ оно может быть не определено. Тогда существует единственная функция f, которая

1) или определена на всем множестве A и согласована с рекурсивным определением, т. е. $f(x) = F(x, f|_{[0,x)})$;

2) или определена на некотором начальном отрезке [0,a) и на нём согласована с рекурсивным определением, причём для точки a и функции f рекурсивное правило неприменимо (отображение F не определено).

Доказательство. Дополним множество *B* специальным элементом $\perp$, который обозначает неопределённость отображения, и модифицируем рекурсивное правило таким образом. Новое правило даёт значение $\perp$ тогда, когда старое правило было не определено, вне зависимости от того, встречалось ли ранее значение $\perp$, или нет.

Применяя к модифицированному правилу теорему 5, получаем некоторую функцию *f'*. Если эта функция не принимает значения $\perp$ на любом из значений своих аргументов, то имеет место случай 1), который приведен в условии теоремы (при *f = f'*). Если же функция *f'* принимает значения $\perp$ при некотором значении *a* своего аргумента, то она принимает это же значение и на всех значениях аргументов больших *a*. Заменяя значение $\perp$ на неопределённость, получаем из функции *f'* функцию *f*. Областью определения функции *f* выступает некоторые начальный отрезок *[0,a)* и, следовательно, имеет место случай 2), приведённый в условии теоремы. □

Теперь можно дать ответ на поставленный выше вопрос. Этот ответ вытекает из теоремы Цермело.

Теорема Цермело. Произвольное множество можно полностью упорядочить.

Доказательство этой теоремы строится на аксиоме выбора и вызывает большое количество нареканий своей неконструктивностью. На счётном множестве полный порядок указать нетрудно, поскольку полный порядок с *N* переносится на данное счётное множество (и этого достаточно для некоторых областей науки о вычислениях). Но в случае множества действительных чисел никакого конкретного полного порядка указать нельзя и, доказав с помощью аксиомы выбора его существование, мы так и не можем себе этот порядок представить.

Пусть *A* некоторое заданное множество. Рассмотрим в каком виде используется аксиома выбора. Допустим, что существует функция φ, определённая на всех собственных подмножествах множества *A*, кроме самого множества *A*, которая ставит в соответствие один элемент за пределами этого подмножества:

$$X \subseteq A \Rightarrow \varphi(X) \in A \setminus X.$$

После того, как такая функция зафиксирована, можно строить полный порядок на множестве *A* таким способом. Наименьшим элементом множества *A* объявляем элемент $a_0 = \varphi(\varnothing)$. За ним идет элемент $a_1 = \varphi(\{a_0\})$. По построению он отличается от a_0. Далее идет элемент $a_2 = \varphi(\{a_0, a_1\})$ и т. д. Если множество *A* бесконечно, то такой процесс можно продолжать и получать последовательность $a_0, a_1, a_2, \ldots$. Если и после этого остаются нерассмотренные элементы в множестве *A*, то строим элемент $w = \varphi(\{a_0, a_1, a_2, \ldots\})$ и так будем продолжать до тех пор, пока не исчерпаем всё множество *A*. Когда этот процесс закончится, то полученный порядок будет полным порядком на множестве *A*. □

Из теоремы Цермело следует, что возможность распространения рекурсивных и индуктивных определений на трансфинитные области было бы конструктивным, если бы была конструктивной аксиома выбора.

Частично упорядоченные множества. Как следует из доказательства теоремы Цермело, построение полного порядка на множестве *A* желает лучшего. В связи с этим, рассмотрим вопрос распространения принципа математической индукции и рекурсии на частично упорядоченные множества (чум). Ответ на этот вопрос даёт такая

Теорема 7. Следующие свойства чум A эквивалентны:

а) любое непустое подмножество множества A имеет минимальный элемент;

б) в множестве А не существует бесконечной строго убывающей последовательности элементов $a_0 > a_1 > a_2 > \ldots$;

в) на множестве А выполняется принцип математической индукции в таком виде: если для каждого x $\in$ A из истинности свойства P(y) для всех y < x вытекает истинность P(x), то свойство P(x) истинно для всех x из A.

Доказательство. **а)** $\to$ **б)**. Пусть произвольное непустое подмножество множества *A* имеет минимальный элемент. Если $a_0 > a_1 > a_2 > \ldots$ - бесконечная строго убывающая последовательность, то подмножество *B* = { $a_0 > a_1 > a_2 > \ldots$ } не будет иметь минимального элемента, поскольку для любого a_n существует меньший элемент a_{n+1}. Получаем противоречие с существованием в множестве *B* минимального элемента.

б) $\to$ **а)**. Пусть *B* - непустое подмножество множества *A*, которое не имеет минимального элемента. Тогда бесконечную убывающую последовательность в множестве *B* можно построить таким образом. Возьмём произвольный элемент $b_0 \in B$. Этот элемент не является минимальным в силу предположения. Возьмём $b_1 \in B$ такой, что $b_0 > b_1$. На том же основании элемент b_1 не является минимальным. Далее находим b_2 < b_1 < b_0 и т. д. В результате получаем бесконечную убывающую последовательность, что противоречит условию б).

а) $\to$ **в)**. Пусть *P(x)* – произвольное свойство элементов из множества *A*. Предположим, что *P(x)* истинное не для всех элементов из *A*. Пусть *B* означает множество элементов, для которых *P(x)* ложно. Свойство *P(* a_0 *)* истинное на минимальном элементе $a_0 \in A$ и поэтому $a_0 \notin B$. Для всех *a* < *b*, де *b* - минимальный элемент в множестве *B* свойство *P(a)* выполняется. Но тогда должно выполняться и *P(b)*. Полученное противоречие доказывает утверждение.

в) $\to$ **а).** Пусть *B* - подмножество множества *A*, в котором отсутствует минимальный элемент. Докажем с помощью индукции, что множество *B* пусто. Пусть *P(x)* означает свойство *x* $\notin$ *B*. Тогда свойство *P(y)* выполняется для всех *y* < *x*, т. е. ни один элемент, меньше *x* не принадлежит к *B*. Если бы *x* принадлежал к *B*, то *x* был бы там минимальным элементом, а это противоречит предположению об отсутствии таких элементов в множестве *B*. □

Доказанная теорема даёт возможность ввести такое определение.

Определение *5.* Чум, для которого выполняется одно из условий выше доказанной теоремы (а, следовательно, и остальные условия), называется *фундированным множеством*.

Примеры фундированных множеств.

1) Множество натуральных чисел N относительно частичного порядка делимости, т. е. *m* < *n* $\Leftrightarrow$ *m* делитель *n*.

2) Множество N $\times$ N относительно частичного порядка (m,n) $\leq$ (m',n') $\Leftrightarrow$ (m < m') $\vee$ (m = m' $\wedge$ $\wedge$ n $\leq$ n'). Для проверки справедливости условия б) выполним такие действия. Пусть

$$(a_0, b_0) \geq \geq (a_1, b_1) \geq (a_2, b_2) \geq \ldots -$$

произвольная последовательность элементов из N $\times$ N. По определению частичного порядка сравниваем первые компоненты пар $a_0 \geq a_1 \geq a_2 \geq \ldots$. В силу фундированности множества N эта последовательность стабилизируется на некотором конечном индексе n. После этого другие компоненты b_i должны убывать и тоже стабилизироваться на некотором конечном индексе m. Поскольку m и n конечны, то это даёт стабилизацию всей последовательности.

3) Рассмотренный случай обобщается на произвольные фундированные множества, что вытекает из следующей теоремы.

Теорема 8. Пусть A и B - два фундированных множества. Тогда их декартово произведение $A\times B$ тоже фундированное множество относительно частичного порядка $(a,b)\le(a',b') \Leftrightarrow [(a < a')\vee(a = a' \wedge b\le b')]$.

Доказательство. В последовательности $(a_0,b_0)\ge(a_1,b_1)\ge(a_2,b_2)\ge\ldots$ сначала стабилизируются первые компоненты, а потом вторые. □

Из этой теоремы вытекает, что декартово произведение $A_1\times A_2\times\ldots\times A_n$ конечного числа фундированных множеств $A_1, A_2,\ldots, A_n$ является фундированным множеством.

4) Сумой *A+B* двух непересекающихся множеств называется их теоретико-множественное объединение. Имеет место

Теорема 9. Сума A+B двух фундированных множеств является фундированным множеством.

Доказательство. Любая последовательность $x_1 > x_2 > \ldots$ либо полностью принадлежит множеству *B*, либо включает элемент $a\in A$ и тогда все последующие элементы принадлежат к множеству *A*. В первом случае справедливость теоремы следует из фундированности множества *B*, а в другому - из фундированности множества *A*. □

Рассмотренные утверждения и примеры фундированных множеств часто стают полезными при доказательстве терминальности циклов в программах. Например, если необходимо доказать, что данный цикл в программе заканчивается, то вводим (если в этом есть необходимость) некоторый натуральный параметр и убеждаемся в том, что он при каждом повторении цикла уменьшается. Тогда, на основе фундированности множества *N*, этот цикл заканчивается через конечное число повторений.

Лемма Цорна и рекурсия. Часто трансфинитную индукцию или рекурсию заменяют ссылкой на лемму Цорна. Рассмотрим лемму Цорна и причини таких ссылок. Для этого напомним некоторые понятия. *Цепью в чум A* называется произвольное линейно упорядоченное его подмножество *B*. Произвольный элемент $a \in A$ (не обязательно должен быть элементом множества *B*), который удовлетворяет условию $(\forall x\in B)\, x\le a$, называется *верхней гранью* подмножества *B* в множестве *A*.

Теорема 10 [Лемма Цорна]. Пусть A - чум, в котором каждая цепь имеет верхнюю грань. Тогда в множестве A существует максимальный элемент. Более того, для любого $a \in A$ существует элемент $a \le b$, который является максимальным в A.

Доказательство. Заметим, что поскольку множество *A* чум, то необходимо различать максимальный и наименьший элементы.

Пусть $a\in A$ – произвольный элемент. Возьмём некоторое пум *I* достаточно большой мощности (большей чем мощность *A*). Построим строго возрастающую функцию $f : I \to A$ при помощи трансфинитной рекурсии. Её значение на минимальном элементе множества *I* положим равным *a*. Допустим, что нам уже известны её значения на всех элементах, меньших некоторого $i \in I$. На основе монотонности *f* эти значения попарно сравнимы между собой. А поэтому существует верхняя грань *s*, которая, в частности, больше или равна *a*. Возьмём некоторый элемент $t > s$ и положим $f(i) = t$. По построению монотонность функции *f* сохраняется. Таким образом, множество *I* равномощно части множества *A*, что противоречит выбору множества *I*. □

Уточним некоторые тонкие моменты этого доказательства, поскольку мы одновременно определяли функцию при помощи трансфинитной рекурсии и доказывали её монотонность при помощи трансфинитной индукции. Наше рекурсивное определение имеет смысл только в тому случае, когда уже построенная часть функции монотонна. Здесь нужно воспользоваться теоремой 6, считая, что следующее значение функции не определено, если уже построенная часть функции не является монотонной, и получить функцию, которая определена на всём множестве *I* или на начальном отрезке. Если функция определена на некотором начальном отрезке, то она монотонна по построению, а поэтому следующее значение тоже определено. А это противоречит частичности функции *f*. □

Ещё одно обстоятельство в доказательстве этой теоремы требует пояснения: можно ли выбрать множество *I*, мощность которого больше мощности *A*? Ответ положительный и вытекает из обобщенной теоремы Кантора.

Теорема 11 [Обобщённая теорема Кантора]. Никакое множество A не может быть равномощно своему булеану B(A). Множество A равномощно некоторому собственному подмножеству множества B(A).

Доказательство. Предположим, что существует взаимно однозначное отображение φ между элементами множества *A* и элементами её булеана *B(A)*. Рассмотрим те элементы $a \in A$, которые не принадлежат соответствующему им подмножеству, и пусть *B* - множество, которое состоит из таких элементов, т. е. $B = \{a \in A : a \notin \varphi(a)$.

Покажем, что множество *B* не соответствует ни одному элементу в множестве *B(A)*. Предположим, что это не так и $\varphi(a) = B$ для некоторого элемента $a \in A$. Тогда

$$a \in B \Leftrightarrow a \notin \varphi(a) \Leftrightarrow a \notin B,$$

на основе построения и предположения, что $\varphi(a) = B$. Полученное противоречие показывает, что множество B не соответствует ни одному элементу. Следовательно, отображение φ не является взаимно однозначным.

Равномощность множества *A* и собственного подмножества *B(A)* очевидна, поскольку соответствие есть таким: элементу $a \in A$ соответствует одноэлементное подмножество *{a}* из *B(A)*. □

Заключение

Проведенный анализ показывает, что рекурсия и индукция распространяется на многие частично упорядоченные и полностью упорядоченные множества. С помощью этих операторов описываются такие абстрактные типы данных как списки, строки, стэки, очереди, очереди с приоритетами, бинарные деревья и т. д. (см. [Крывый-10, Hein-95]. Из приведенного анализа также следует, что рекурсию и индукцию можно было бы распространить на произвольные множества, если бы аксиома выбора была конструктивной. А точнее было известно как вычислять функцию, фигурирующую в этой аксиоме.

Библиография

[Ахо-79] Ахо А., Хопкрофт Дж., Ульман Дж. Структуры данных и алгоритмы. –М.: Издат. Дом «Вильямс». - 2000. - 382 с.

[Гудст-70] Гудстейн Р.Л. Рекурсивный математический анализ. – М.: Наука. - 1970. - 472 с.

[Крывый-10] Крывый С.Л. Астрактные типы данных как многоосновные алгебраические системы. – К.: ж. «Инженерия программного обеспечения». – 2010. - № 3. – С. 3 -18.

[Hein-95] Hein J.L. Discrete Mathematics. – Sudbury, Massachusetts: Jones and Bartlett Publishers. -1995.-656 p.

[Мальцев-70] Мальцев А.И. Алгебраические системы . – М.: Наука. - 1970. - 392 с.

Сведения об авторе

Крывый Сергей *– Киевский национальный университет им. Тараса Шевченка, Украина, Киев, 03680, просп. Глушкова, 4д, Факультет кибернетики. e-mail: krivoi@i.com.ua*

ПРИНЦИПЫ ПОСТРОЕНИЯ ИНТЕГРИРОВАННЫХ СИСТЕМ МУЛЬТИ-АГЕНТНОЙ НАВИГАЦИИ И ИНТЕЛЛЕКТУАЛЬНОГО УПРАВЛЕНИЯ МЕХАТРОННЫМИ РОБОТАМИ

Тимофеев Адиль Васильевич, Юсупов Рафаэль Мидхатович

Abstract: Обсуждаются проблемы мульти-агентной (групповой) навигации и интеллектуального управления движением мехатронных роботов в динамической среде с препятствиями. Определяются понятия глобальной управляемости и стабилизируемости программных движений мульти-агентных робототехнических систем и рассматриваются критерии оптимизации, устойчивости и синхронизации программных движений мехатронных роботов-агентов. Даётся сравнительный анализ четырёх поколений локальных систем навигации и управления движением и стратегий централизованного, децентрализованного и мульти-агентного управления информационными потоками в сложных робототехнических системах. Значительное внимание уделяется методам мульти-агентной навигации и интеллектуального управления и их приложениям в информатике и робототехнике.

Keywords: Интегрированные системы, интеллектуальное управление, групповая навигация, мехатронные агенты, мульти-агентные робототехнические системы.

ACM Classification Keywords: H.1.1 Systems and Information Theory

Введение

В последние годы в России и за рубежом значительное внимание уделяется исследованию проблем интеллектуализации локальных систем управления роботов и разработке стратегий групповой навигации и управления движением сложных робототехнических систем (РТС) т.е. группы (коллектива) роботов, объединенных общей (глобальной) целью. В связи с бурным развитием систем навигации и управления движением мехатронных роботов как подвижных агентов с элементами искусственного интеллекта и необходимости организации их коллективной работы возникла потребность создания основ теории мульти-агентных робототехнических систем (МАРС). Также МАРС объединяют группу агентов-роботов (например, мобильных роботов или беспилотных летательных аппаратов) для достижения общей (глобальной) цели в динамической среде с препятствиями или запретными зонами и возможным противодействием.

Фундаментальные и поисковые исследования в области адаптивного и интеллектуального управления роботами и робототехническими системами (РТС) активно проводились в России в ряде научных школ (МГТУ, СПИИРАН, СПбГУ, МГУ, МИРЭА, НИИ МВС, ЦНИИ РТК, ИПМ РАН и др.), начиная с 70-х годов XX века [1–17]. Важная роль в становлении и развитии этих исследований принадлежит академикам И.М.Макарову, Д.Е.Охоцимскому, Е.П.Попову, Г.С.Поспелову и Ф.Л.Черноусько, членам-корреспондентам РАН В.А.Якубовичу, В.А.Лопоте, Г.Г.Сербрякову, Е. Д.Теряеву, Р.М.Юсупову и И.А.Каляеву и профессорам Е.И.Юревичу, В.М.Лохину, А.В.Тимофееву и др.

Проблемы и методы адаптивной навигации и интеллектуального управления движением роботов, а также задачи группового управления РТС обсуждались на ряде Международных и Российских конференций, проведённых в последние десятилетия в России, в том числе на 2-х Международных конференциях “Мехатроника и робототехника”, 20-и конференциях “Экстремальная робототехника”, 7-и конференциях

“Мехатроника, автоматизация, управление”, 6-и конференциях “Управление и информационные технологии” и 3-ех Мультиконференциях по проблемам управления.

Параллельно в России проводились поисковые исследования и опытно-конструкторские разработки по созданию интегрированных систем навигации и управления движением объектов различного типа и назначения (в основном для экстремальных сред и частично неопределённых условий эксплуатации подвижных объектов и роботов). Достижения специалистов из России и зарубежный опыт в этой области обсуждались на 17-ти Международных конференциях по интегрированным навигационным системам, организованных ЦНИИ «Электроприбор», и на 15-ти научных сессиях Международной академии навигации и управления движением, президентом которой является академик В.Г.Пешехонов.

Первый международный проект по исследованию МАРС на тему “Multi-Agent Robot Systems for Industrial Applications in the Transport Domain” был выполнен в 1997–1999 годах по Европейской программе COPERNICUS. Координатором этого проекта был Prof., Dr.–Ing. U.Rembold (University of Karlsruhe, Department of Computer Science, Institute for Process Control and Robotics, Germany), а его участниками – Institute for Informatics and Automation of Russian Academy of Sciences (Russia),Unite de Recherche INRIA Rhone–Alpes (France), Technical University of Budapest (Hungary), Technical University of Poznan (Poland), Belorussian State University, (Belarus), Ufa State Aviation Technical University (Russia), Daimler-Benz AG (Germany), Beta Computer Automation GMBH (Germany).

Развиваемые в настоящей работе новые информационные и телекоммуникационные технологии и интеллектуальные системы навигации и управления движением для МАРС в значительной степени базируются на оригинальных научных результатах поисковых и ориентированных фундаментальных исследований, полученных в последние годы при поддержке грантов Российского фонда фундаментальных исследований (РФФИ), программ Президиума РАН и государственных заказов Министерства образования и науки РФ.

1. **Глобальная управляемость и оптимизация программных движений робототехнических систем**

Рассмотрим некоторую робототехническую систему (РТС) S, состоящую из n роботов $r_i, i = 1,2,, \dots, m.$ Обычно роботы и РТС функционируют в динамической среде E с препятствиями или запретными зонами $O_j \in E, j = 1,2,, \dots, q.$ В роли динамических препятствий для одних роботов r_i из РТС S могут выступать другие роботы $r_i, j \neq i.$

РТС будем называть гомогенной, если она состоит из однотипных роботов (например, только из манипуляционных роботов). Сложную РТС будем называть гетерогенной, если её структура включает в себя несколько разнотипных роботов (например, манипуляционных и мобильных роботов) или разных гомогенных РТС меньшего масштаба, т.е. гомогенных подсистем.

Архитектура гомогенных и гетерогенных РТС по существу является сетевой и территориально распределённой, т.е. включает в себя множество роботов, взаимодействующих между собой с помощью физических, сенсорных (информационных), управляющих и коммуникационных каналов прямой и обратной связи. Поэтому возникает потребность в разработке новых принципов сетевой организации, управления движением, навигации, обработки и передачи информации в сложных РТК в динамической окружающей среде с препятствиями или запретными зонами.

Необходимость в сетевом управлении и групповой навигации возникла прежде всего в глобальных инфотелекоммуникационных сетях (например, в Internet), в робототехнике и гибких автоматизированных производственных системах, а также в сложных автоматизированных системах вооружений (например, в стереоцентрических и мульти-агентных сетях оборонного назначения).

В сложных гомогенных и гетерогенных РТС под действием вектора управляющих воздействий

$$u(t) = |u_i(t)|_{i=1}^{n}, t \in [t_0, t_T], \quad (1)$$

каждый робот $r_i(u_i)$ может выполнить некоторый набор локальных технологических операций $l_j(r_i), j = 1,2,\dots,p, p \geq m$, на заданном интервале времени $T = t_T - t_0$. В результате выполнения этих локально управляемых операций изменяются состояния как самой РТС *S(t)*, так и окружающей её среды *E(t)*.

Предположим, что эволюцию РТС *S(t)* и окружающей среды *E(t)* под действием вектора управления u(t) формально можно описать дифференциальными уравнениями вида

$$\dot{S} = F(S, u, E), \quad S(t_0) = S_{0,} \tag{2}$$

$$\dot{E} = \Phi(S, E), \quad E(t_0) = E_{0,} \tag{3}$$

где F и Ф – некоторые операторы, а S_0 и E_0 – начальные состояния РТС и среды.

Общим (комплексным) динамическим состоянием РТС S и окружающую её среды E в текущий момент времени t будем называть вектор-функцию вида

$$x(t) = \begin{vmatrix} S(t) \\ E(t) \end{vmatrix}, t \in [t_0, t_T]\,. \tag{4}$$

Обычно на вектор группового управления (1) и состояний РТС и среды (4) наложены ограничения вида

$$u(t) \in Q_u \in R^n \tag{5}$$

$$f\big(x(t)\big) \leq 0, t \in\ [t_0, t_T]\,. \tag{6}$$

Глобальная (общая) цель управления РТС заключается в том, чтобы синтезировать такой закон группового управления локальными технологическими операциями роботов $L = |l_i(r)|_{j=1}^{p}$, чтобы перевести РТС *S* и окружающую её среду *E* из заданного начального состояния в желаемое конечное (целевое) состояние, т.е. в процессе групповой навигации и управления (5) должны быть выполнены граничные условия вида

$$x(t_0) = x_0,\ \ x(t_T) = x_T, \tag{7}$$

с учётом заданных ограничений на вектор управлений (5) и вектор состояний (6).

РТС (2) будем называть глобально управляемой в динамической среде (3), если существует закон группового управления (вообще говоря, зависящий от текущего стояния РТС и среды) вида

$$u_p(t) = U\big(t, x(t)\big) \in Q_u\,, t \in\ [t_0, t_T]\,, \tag{8}$$

обеспечивающий выполнение граничных условий (7) с учетом ограничений (5) и (6). Соответствующее этому групповому уравнению (8) движение $x_p(t), t \in\ [t_0, t_T]$, будем называть программным движением (ПД) РТС *S(t)* в динамической среде *E(t) с* препятствиями или запретными зонами.

Среди множества групповых управлений (8) и ПД РТС можно выделить наилучшие (оптимальные) управление и соответствующее ему ПД. Для этого зададим некоторый интегральный функционал качества групповых ПД вида

$$K\big(x_p\big) = \int_{t_0}^{t_T} \psi\big(x_p(t), \dot{x}_p(t)\,\big)dt \tag{9}$$

Тогда оптимальным ПД $x_p^{opt}(t), t \in [t_0, t_T]$, и соответствующим ему оптимальным групповым управлением $u_p^{opt}(t)$ будем называть то ПД и закон управления РТС, при которых достигается экстремум функционала качества (9), т.е.

$K(x_p^{opt}) = \min K\,\big(x_p\big), t \in [t_0, t_T]$.

Следует отметить, что критерии глобальной управляемости мехатронных роботов с нелинейной динамикой сформулированы в [2,6,10]. Методы аналитического синтеза и оптимизации ПД мехатронных роботов, обеспечивающие выполнение граничных условий (7) с учётом заданных ограничений на векторы управлений (7) и состояний (8) при наличии препятствий предложены в [2,6,7,18].

2. Устойчивость, стабилизация, декомпозиция и синхронизация программных движений роботов

Предположим, что программное движение (ПД) РТС $x_p\ (t),\ t \in [t_0, t_T]$, и соответствующий ему закон программного управления (8), удовлетворяющие граничным условиям (7) и ограничениям (5) и (6), существуют. Тогда ПД называется практически (или асимптотически) устойчивым по отношению к начальным возмущениям $e(t_0) = x_0 - x_p(t_0)$ (или, возможно, к другим возмущениям), если существует закон группового управления роботами (8) такой, что в замкнутой этим управлением РТС и окружающей её динамической среде выполняются условия вида

$$\|e(t)\| = \|x_p(t) - x(t,u)\| \leq \varepsilon,\ \ t \geq t_p \geq t_0, \tag{10}$$

где $\varepsilon \geq 0$ – параметр, определяющий желаемую точность осуществления ПД, а $T_p = t_p - t_0$ – время переходного процесса. Очевидно, что для достижения цели управления (7) должно выполняться условие

$$t_p \leq t_T \tag{11}$$

Закон группового управления РТС (8), обеспечивающий выполнение целевых условий (10), (11), будем называть стабилизирующим. Этот закон управления является декомпозирующим, если целевые условия выполняются независимо для каждого робота из РТС, т.е. $e_i(t)$ не зависит от $e_j(t), j \neq i$. Это означает, что локальное управление каждым роботом r_i осуществляется независимо от локального управления другими роботами $r_j, j \neq i$, с компенсацией возможных перекрёстных динамических связей между роботами РТС.

В некоторых случаях (например, на конвейерах) от закона группового управления требуется обеспечить полную или частичную синхронизацию движений различных роботов из РТС. В этих случаях целевые условия (10) приобретают вид

$$\|e_{ij}(t)\| = \|x_{p,i}(t) - x_j(t,u_j)\| \leq \varepsilon_{ij},\ \ i \neq j, \qquad t \geq t_{p,i,j}\ . \tag{12}$$

Возникают также задачи группового управления роботами, связанного с достижением консенсуса (concensus), когда каждый робот из РТС стремится, чтобы его ПД было близко с ПД своих соседей или рандеву (rendezvous), когда все роботы и РТС должны попасть в заданное состояние в заданный момент $T = t_T - t_0$.

Следует отметить, что для локальных систем навигации и управления движением мехатронных роботов с нелинейной динамикой законы стабилизирующего, модального (спектрального) и декомпозирующего управления были предложены в [2,6,18].

3. Четыре поколения систем навигации и управления движением роботов

В работах [1,2] была предложена классификация локальных систем навигации и управления движением (СНУД) роботов на четыре класса:

1) программные СНУД;

2) адаптивные СНУД;

3) интеллектуальные СНУД;

4) нейросетевые СНУД.

Программные СНУД роботов основаны на классических принципах программного или оптимального управления движением, если модель динамики роботов или РТС (2) полностью известна, а среда (или связанные с ней возмущения) известны и неизменны, т.е. в модели динамики среды

$\dot{E} = 0,$ а $E = S, t \in [t_0, t_T]$ (см., например, [2]).

Адаптивные СНУД базируются на современных принципах робастного или адаптивного управления, когда модель динамики РТС (2) и среды E (или связанных с ней возмущений) (3) частично неизвестны. В этом случае программные СНУД дополняются специальными средствами адаптации или идентификации факторов неопределённости или нестационарности, к числу которых можно отнести неизвестные возмущения или препятствия [2,6].

Интеллектуальные СНУД дополняются не только средствами адаптации, но и некоторыми элементами (алгоритмами) искусственного интеллекта (например, способностью распознавать речевые команды или идентифицировать неизвестные препятствия). Как правило, эти элементы искусственного интеллекта описываются логическими, алгебраическими и нечёткими алгоритмами. Поэтому они могут быть программно реализованы на традиционных микропроцессорах и компьютерах [3–5,6].

Нейросетевые СНУД основываются на обучении, самоорганизации и распараллеливании процессов обработки информации, навигации и управления на нейронных сетях или нейрокомпьютерах [7]. В этом заключается их основное отличие и преимущество по сравнению с СНУД роботов первых трёх поколений.

4. **Стратегии централизованного, децентрализованного и мульти-агентного управления робототехническими системами**

Важно отметить, что роботы редко используются изолированно. Обычно они входят в состав РТС и предназначены для группового (коллективного) выполнения некоторой сложной общей (глобальной) задачи, которую каждый робот самостоятельно (автономно) решить не может.

Сложная гетерогенная РТС может включать в себя десятки или сотни роботов или гомогенных РТС, функционирующих в динамической среде с препятствиями. В связи с этим важное значение приобретают проблемы системного (сетевого) анализа и синтеза интегрированных систем групповой навигации, управления движением и обработки сенсорной информации. Однако проектирование и создание интегрированных систем невозможно без организации и координации взаимодействия роботов в РТС с помощью коммуникационной сети [5,13,14].

Математической моделью такой, вообще говоря, динамической коммуникационной сети РТС может служить коммуникационный граф вида

$$G(t) = \{R(t), C(t), w(t)\}, t \in [t_0, t_T] \qquad (13)$$

где *R* – множество узлов, соответствующих локальным системам навигации, управления движением и обработки информации роботов $r_i, i = 1,2,\ldots,m$, *C* – множество каналов связи между ними, а W – множество весов (например, длина или пропускная способность) каналов связи.

В процессе решения общей (глобальной) задачи структура и параметры РТС могут изменяться (например, некоторые роботы или каналы связи между ними могут выйти из строя). Тогда будут изменяться структура (топология узлов и каналов связи) или параметры (веса) коммуникационного графа (13). Этот граф характеризует информационное взаимодействие между роботами r_i из РТС $S(t)$. Поэтому его структура тесно связана со стратегией группового управления ПД роботов и РТС (8), обеспечивающей достижение общей (глобальной) цели управления.

Первоначально (начиная с 70-х годов XX века) стратегии группового управления роботами в РТС разделялись на два класса:

– стратегии централизованного (глобального) управления РТС из общего командного центра;

– стратегии децентрализованного управления, основанные на локальном управлении каждым роботом из РТС.

Преимущества стратегии централизованного (глобального) управления РТС заключаются в простоте иерархической организации и программно-аппаратной реализации систем группового (в том числе оптимального) управления. Однако её недостатки проявляются в низкой живучести (надежности) и возможности искажений и временных задержках при передаче команд программного управления от верхнего уровня иерархии к нижнему.

Преимущества стратегии децентрализованного (локального) управления роботами из РТС заключаются в высокой локальной производительности (малом времени принятия локальных решений) и параллелизме локального управления роботами, а также в повышенной живучести (надёжности) РТС. В тоже время ей свойственны такие недостатки, как повышенные требования к надёжности узлов и каналов информационной связи между роботами и невысокое качество группового управления РТС, т.к. даже из локальной оптимальности управления отдельными роботами, вообще говоря, не следует глобальная оптимальность управления РТС в целом.

Сравнительный анализ преимуществ и недостатков традиционных стратегий централизованного и децентрализованного управления территориально распределённых РТС привёл к необходимости разработки (начиная с 90-х годов XX века) новой гибридной (смешанной) стратегии мульти-агентного управления РТС на современном уровне развития робототехники, мехатроники и инфотелекоммуникационных технологий [5–17]. Этот уровень развития характеризуется всё более широким внедрением роботов с интеллектуальным и нейросетевым управлением, глобальных систем навигации типа ГЛОНАСС и (или) GPS и инфотелекоммуникационных систем типа Internet и Grid.

5. **Принципы интеграции систем навигации и управления движением в мульти-агентных робототехнических системах**

В современной робототехнике и мехатронике, а также в информационных технологиях навигации и теории управления движением, существует ряд проблем, связанных с проектированием, созданием и интеграцией систем адаптивной (локальной) и мульти-агентной (групповой) навигации, функциональной диагностики и интеллектуального управления движением мобильных роботов в динамической среде с известными или неизвестными препятствиями. Некоторые из этих проблем, а также перспективные подходы и новые методы их решения рассмотрены в [1–20].

В настоящей работе обсуждаются принципы проектирования интегрированных систем мульти-агентной навигации и интеллектуального управления мобильными роботами и РТС.

Мобильный робот рассматривается как интеллектуальный подвижный мехатронный агент, имеющий на борту [1,2]:

– сенсорную систему датчиков внутренней и внешней информации;

– интеллектуальную систему навигации и управления движением;

– двигательную систему с традиционным (например, гусеничным или четырёхколёсным) или нетрадиционным (например, с одноосным двухколёсным или ползающим, т.е. перемещающимся по поверхностям произвольной ориентации) шасси;

– коммуникационную систему для информационного взаимодействия с человеком-оператором и другими роботами-агентами.

Важными задачами при создании мобильных мехатронных роботов как подвижных агентов является интеллектуализация бортовых интегрированных систем навигации и управления движением, обеспечивающая возможность автоматического планирования и оптимизации движений для обхода известных или неизвестных препятствий, адаптации к динамическим факторам неопределённости (например, к возмущениям и неизвестным подвижным препятствиям) и к возможным дефектам или отказам, распознавания образов (например, стерео-изображений препятствий), моделирования и анализа сложных 3-D сцен и диагностики состояний мобильного робота-агента в реальном масштабе времени [1–7, 18–20].

Новые актуальные проблемы возникают при групповом (коллективном) использовании мобильных мехатронных роботов-агентов в составе МАРС для решения общей (глобальной) задачи.

Группа (коллектив) мобильных мехатронных роботов-агентов, объединённых общей (глобальной) целью и функционирующая в едином информационном и коммуникационном пространстве на базе соответствующих стандартов, команд и форматов передачи потоков данных, называется мобильной мульти-агентной робототехнической системой (МАРС).

Ключевыми проблемами при создании интегрированных систем навигации и управления движением мобильных МАРС являются [7–10]:

– декомпозиция общей (глобальной) задачи, возлагаемой на группу (коллектив) мобильных роботов, на ряд локальных задач для каждого робота-агента,

– мульти-агентное (групповое) планирование маршрутов движения и навигация мобильных роботов-агентов в динамической среде с известными или неизвестными препятствиями,

– функциональная диагностика и отказоустойчивое управление движением мобильных роботов-агентов,

– интеллектуализация и интеграция систем навигации и управления движением на базе бортовых средств распознавания сложных образов (мульти-изображений, речевых команд и т.п.) и мобильных навигаторов (спутниковые системы ГЛОНАСС и GPS и т.п.) в составе мобильных МАРС.

Важными задачами являются также организация интеллектуального человеко-машинного интерфейса и комплексирование перспективных решений этих проблем для создания интегрированных систем мульти-агентной навигации и интеллектуального управления движением мобильных роботов как агентов МАРС.

Заключение

Актуальность проведённых исследований связана с повышенным интересом во всём мире к интеллектуальным мехатронным роботам как подвижным агентам и мобильным МАРС ввиду их способности автоматически функционировать под контролем человека как в традиционных отраслях производства (машиностроение, приборостроение, микроэлектроника и т.п.), так и в нетрадиционных областях и экстремальных средах (космическая и подводная робототехника, атомная энергетика, ликвидация техногенных аварий и террористических угроз и т.п.). Следует отметить, что круг задач, решаемых мехатронными роботами-агентами и мобильными МАРС, непрерывно расширяется и усложняется (необходимость эффективно работать в экстремальных средах в условиях неопределённости или противодействия при наличии препятствий или запретных зон и т.п.).

Большую научную значимость и актуальность имеют поисковые исследования в области создания интегрированных систем мульти-агентной (групповой) навигации и интеллектуального управления движением мехатронных роботов-агентов и мобильных МАРС в динамических (изменяющихся) средах с препятствиями. Сегодня этим инновационным исследованиям и опытно-конструкторским разработкам во всём мире уделяется приоритетное внимание.

Работа выполнена при частичной поддержке грантов РФФИ № 09–08–00767–а "Моделирование поведения и анимация движений подвижных агентов" и РФФИ–ГФЕН Китая № 10–08–91159–ГФЕН–а “Исследование научных проблем интеллектуального управления магистральными транспортными средствами и мобильными роботами” и Программы № 13 (Грид) Президиума РАН.

Литература

[1] Тимофеев А.В. Роботы и искусственный интеллект. – М.: Наука, 1978.
[2] Тимофеев А.В. Адаптивные робототехнические комплексы. Л.: Машиностроение, 1988.
[3] Тимофеев А.В., Юсупов Р.М. Интеллектуализация систем автоматического управления. – Известия АН. Техническая кибернетика, 1994, № 5.
[4] Timofeev A. V., Yusupov R.M. Principles of Artificial Intelligence Applied to Adaptive Control Systems – Proceedings of International Conference on CAD/CAM, Robotics and Factories of Future (Saint-Petersburg, Russia, SPIIRAS, 1993), vol. 2, 1993, pp. 415-421.

[5] Тимофеев А.В., Сырцев А.В. Модели и методы маршрутизации потоков данных в телекоммуникационных системах с изменяющейся динамикой. М.: Новые технологии, 2005.

[6] Тимофеев А.В. Методы высококачественного управления, интеллектуализации и функциональной диагностики автоматических систем. Часть I, Часть II. – Мехатроника, автоматизация, управление, 2003, № 5, 2004, № 2.

[7] Timofeev A. V. Intelligent Control Applied to Non-Linear Systems and Neural Networks with Adaptive Architecture //International Journal on Intelligent Control, Neurocomputing : Fuzzy Logic. 1996. P. 1–18.

[8] Тимофеев А. В., Сырцев А. В. Мультиагентная и нейросетевая маршрутизация потоков данных в телекоммуникационных сетях // Труды 10-й международной конференции "Knowledge–Dialogue–Solution" (16–26 июня. 2003, Варна) 2003. С. 187–190.

[9] Тимофеев А. В. Модели мульти-агентного диалога и информационного управления в глобальных телекоммуникационных сетях // Труды 10-й международной конференции "Knowledge–Dialogue–Solution" (16-26 июня, 2003, Варна). 2003. С. 180–186.

[10] Cai Z., He H., Timofeev A. Navigation and Control of Mobile Robots in Unknown Environment: A Survey// Proceedings of 10-th International Conference on Integrated Navigation Systems (June 27–29, St-Petersburg). 2003. Vol. 1. P. 158–166.

[11] Timofeev A. V. Models for Multi-Agent Dialogue and Informational Control in Global Telecommunicational Networks // International Journal "Information Theories and Their Applications". 2003. N 1.

[12] Timofeev A. V., Syrtzev A. V. Neural Approach in Multi-Agent Routing for Telecommunicational Networks // International Journal "Information Theories and Their Applications". 2003. N 10. P. 167–172.

[13] Timofeev A. V. Multi-Agent Information Processing and Adaptive Control in Global Telecommunication and Computer Networks // International Journal "Information Theories and Their Applications". 2003. N 10. P. 54–60.

[14] Timofeev A. V. Adaptive Control and Multi-Agent Interface for Infotelecommunication Systems of New Generation // International Journal "Information Theories & Applications". 2004. Vol. 11.

[15] Охтилев М.Ю., Соколов Б.В., Юсупов Р.М. Интеллектуальные технологии управления структурной динамикой сложных технических объектов. М.: Наука, 2006.

[16] Макаров И.М., Лохин В.М., Манько С.В., Романов М.П. Искусственный интеллект и интеллектуальные системы управления. М.: Наука, 2006.

[17] Каляев И.А., Гайдук А.Р., Капустян С.Г. Модели и алгоритмы коллективного управления в группах роботов. М.: Наука, 2009.

[18] Зотов Ю.К., Тимофеев А.В., Шишкин Д.С. Информационные технологии навигации и управления полетом малоразмерных летательных роботов корабельного базирования. – Информационно-измерительные и управляющие системы, № 8, т. 6, 2008.

[19] Лютикова Л.А, Тимофеев А.В., Сгурев В.В., Йоцов В.И. Развитие и применение многозначных логик и сетевых потоков в интеллектуальных системах. // Труды СПИИРАН, вып. 2, 2005. С. 114–126.

[20] Тимофеев А.В., Дерин О.А., Гуленко И.Е., Андреев В.А. Распознавание объектов в сложных мультиизображениях и методы и средства видеозахвата движений. - Мехатроника, автоматизация, управление, № 6 (111), 2010.

Сведения об авторах

***Юсупов Рафаэль Мидхатович**, Директор Санкт-Петербургского института информатики и автоматизации РАН (с 1991 г.), 1934 года рождения, член-корреспондент Российской Академии наук (2006), доктор технических наук (1968 г.), профессор (1974 г.), Заслуженный деятель науки и техники РФ (1984 г.).Тел.: (812)328-3311, (812)328-34-11 Факс: (812)328-4450 E-mail: yusupov@iias.spb.su*

***Тимофеев Адиль Васильевич**, Заведующий лабораторией информационных технологий в управлении и робототехнике Санкт-Петербургского института информатики и автоматизации РАН (с 1990 г.), доктор технических наук (1984 г.), профессор (1986 г.), Заслуженный деятель науки и техники РФ (2002 г.).Тел.: (812)328-0421 Факс: (812)328-4450 E-mail: tav@iias.spb.su*

СИММЕТРИЯ В ЗАПИСИ ГЕНЕТИЧЕСКОЙ ИНФОРМАЦИИ В ДНК

Анатолий Гупал, Александра Вагис

***Аннотация**: Показано, что для нитей ДНК возможно два вида симметрии, но в природе реализован один более эффективный способ записи и считывания информации. Доказано, что из симметрии последовательностей оснований вытекает симметрия коротких последовательностей, в том числе отдельных оснований. На основе модели цепей Маркова показано, что симметрия последовательностей оснований вытекает из симметрии пар оснований.*

***Ключевые слова:** основания, комплементарность, симметрия, цепь Маркова, переходные вероятности.*

ACM Classification Keywords: G.3 Probability and statistics.

Введение

Симметрия в записи оснований, подсчитанных по нитям в хромосомах ДНК, исследовалась в работах [1, 2]. Соотношения симметрии приведены в виде коротких формул, что значительно упрощает восприятие этих результатов и является основой построения математического аппарата для получения новых результатов. Статистический анализ подтвердил выполнение соотношений симметрии на геномах бактерий, растений, высших организмов (примерно сто геномов), в том числе и на ДНК человека [1, 2]. Таким образом, в записи генетической информации в ДНК явно наблюдается симметрия, однако до настоящего времени не выяснены причины, которые объясняют этот феномен в природе.

В [3] получены новые правила в записи оснований по одной нити в хромосомах ДНК. Доказано, что из симметрии последовательностей оснований вытекает симметрия коротких последовательностей, в том числе отдельных оснований. Выведены новые связывающие ограничения для пар и троек оснований. На основе модели цепей Маркова показано, что симметрия для троек и коротких последовательностей оснований вытекает из симметрии пар оснований. В настоящей работе исследованы свойства двух видов симметрии для противоположной и одинаковой полярности цепей ДНК.

1. Противоположная полярность цепей ДНК

ДНК имеет форму двойной спирали, информация записана в четырехбуквенном алфавите оснований: аденин (A), цитозин (C), гуанин (G), тимин (T). Известно, что C – G, A – T – комплементарные пары оснований, связывающие две цепи. Хромосомы – неделимые участки ДНК, в них содержится информация относительно тысяч генов, поэтому расчеты проводились на уровне всей хромосомы, а не на уровне отдельного гена. Запись и считывание оснований у первой нити хромосомы ДНК выполняется слева направо в направлении $5' \to 3'$ $5' \to 3'$, а у второй комплементарной нити в направлении $5' \to 3'$ справа налево (рис.1, модель Уотсона-Крика). Приводимые ниже соотношения, как правило, выполняются приближенно.

$$
\begin{array}{ll}
5' \_\_\_\_\_\_\_\_ \mathrm{AC} \_\_\_\_\_\_\_\_ \mathrm{GT} \_\_\_ \rightarrow \_\_ 3' & \text{1-st string} \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \updownarrow\updownarrow \ldots\ldots\ldots\ldots & \\
3' \_\_\_\_\_\_\_\_ \mathrm{TG} \_\_\_\_\_\_\_\_ \mathrm{CA} \_\_\_\_ \leftarrow \_\_ 5' & \text{2-nd string}
\end{array}
$$

Рис 1. Модель Уотсона-Крика

Симметрия оснований. Для оснований, записанных по одной нити ДНК хромосомы, выполняются приближенные соотношения

$$n(\mathrm{A}) = n(\mathrm{T})\,,\ n(\mathrm{C}) = n(\mathrm{G}) \qquad (1)$$

где $n(i)$ – количество оснований i, $i \in \{\mathrm{A,C,G,T}\}$, вычисленных на одной нити. Из соотношений (1) вытекает, что количество каждого основания, подсчитанного по первой и второй нити, совпадает:

$$\begin{aligned} &n(\mathrm{A},1) = n(\mathrm{A},2)\,,\ n(\mathrm{T},1) = n(\mathrm{T},2)\,,\\ &n(\mathrm{C},1) = n(\mathrm{C},2)\,,\ n(\mathrm{G},1) = n(\mathrm{G},2)\end{aligned} \qquad (2)$$

Таким образом, имеет место симметрия относительно записи оснований по каждой нити ДНК. Отсюда следует важный вывод о том, что веса двух нитей совпадают.

Симметрия пар оснований. Расчеты показали, что для пар оснований выполняются соотношения

$$\begin{aligned} &n(\mathrm{AC}) = n(\mathrm{GT})\,,\ n(\mathrm{AG}) = n(\mathrm{CT})\,, n(\mathrm{TC}) = n(\mathrm{GA})\,,\\ &n(\mathrm{TG}) = n(\mathrm{CA})\,,\ n(\mathrm{AA}) = n(\mathrm{TT})\,,\ n(\mathrm{CC}) = n(\mathrm{GG})\end{aligned} \qquad (3)$$

или, короче, в виде формулы

$$n(ij) = n(\overline{j}\,\overline{i}) \qquad (4)$$

где $i, j \in \{\mathrm{A,C,G,T}\}$, $\overline{\mathrm{A}} = \mathrm{T}$, $\overline{\mathrm{C}} = \mathrm{G}$, $\overline{\mathrm{T}} = \mathrm{A}$, $\overline{\mathrm{G}} = \mathrm{C}$. Заметим, что пары AT, TA , CG GC не присутствуют в (3), поскольку они приводят к тавтологии.

Из соотношений (3), (4) вытекает симметрия относительно записи 16 пар оснований по каждой нити ДНК:

$$n(ij,1) = n(ij,2) \qquad (5)$$

где $i, j \in \{\mathrm{A,C,G,T}\}$. Известно, что соотношения

$$\hat{p}(ij) = \frac{n(ij)}{n(i)} \qquad (6)$$

где $n(ij)$ – число пар (ij), $i, j \in \{\mathrm{A,C,G,T}\}$, $n(i)$ – число оснований i в цепи хромосомы, представляют собой оценки переходных вероятностей для однородных цепей Маркова. Из (5) и (6) вытекает, что вторая комплементарная нить в направлении $5' \rightarrow 3'$ имеет такие же оценки переходных вероятностей $\hat{p}(ij)$, как и исходная первая нить (рис.1).

Легко заметить, что для любой последовательности без пропусков букв с точностью до единицы выполняются соотношения

$$\begin{aligned} n(i) &= n(\mathrm{A}i) + n(\mathrm{C}i) + n(\mathrm{G}i) + n(\mathrm{T}i) =\\ &= n(i\mathrm{A}) + n(i\mathrm{C}) + n(i\mathrm{G}) + n(i\mathrm{T}),\end{aligned} \qquad (7)$$

где $i \in \{\mathrm{A,C,G,T}\}$, то есть количество каждой буквы текста можно подсчитать на основе количеств пар букв. Для основания A из (7) получаем связывающее ограничение для пар AT, TA, которые не входят в (3),

$$n(\mathrm{CA}) + n(\mathrm{GA}) + n(\mathrm{TA}) = n(\mathrm{AC}) + n(\mathrm{AG}) + n(\mathrm{AT}) \quad (8)$$

для основания C из (7) – ограничение для пар CG и GC

$$n(\mathrm{AC}) + n(\mathrm{GC}) + n(\mathrm{TC}) = n(\mathrm{CA}) + n(\mathrm{CG}) + n(\mathrm{CT}) \quad (9)$$

Утверждение 1. Из симметрии пар оснований вытекает симметрия оснований.

Симметрия троек оснований. Кодоны (тройки оснований) связаны следующими соотношениями:

$$n(ijk) = n(\overline{k}\,\overline{j}\,\overline{i}) \quad (10)$$

Здесь $n(ijk)$ – число троек оснований (ijk), $i, j, k \in \{\mathrm{A,C,G,T}\}$, $(\overline{k}\overline{j}\overline{i})$ – антикодон кодона (ijk). Аналогично (5) из соотношений (10) вытекает симметрия относительно записи 64 троек оснований для каждой нити ДНК:

$$n(ijk,1) = n(ijk,2) \quad (11)$$

По аналогии с (8), (9), используя соотношения (10), для пар (3) выводятся шесть связывающих ограничений для троек оснований.

Утверждение 2. Из симметрии троек оснований вытекает симметрия пар оснований.

Поскольку симметрия в записи оснований по нитям в ДНК обнаружена эмпирически и в настоящее время не существует объяснения этого феномена в природе, важно построить модель, которая будет подтверждать симметрию последовательностей оснований на основе симметрии коротких последовательностей.

Утверждение 3. Для модели однородной цепи Маркова симметрия троек оснований вытекает из симметрии оснований и симметрии пар оснований.

Из соотношений (1), (4) следует, что для однородной цепи Маркова оценки вероятностей троек оснований (ijk) и $(\overline{k}\overline{j}\overline{i})$ совпадают

$$n\hat{p}(ijk) = \frac{n(i)n(ij)n(jk)}{n(i)n(j)} = n\hat{p}(\overline{k}\overline{j}\overline{i}) = \frac{n(\overline{k})n(\overline{k}\overline{j})n\left(\overline{j}\overline{i}\right)}{n(\overline{k})n(\overline{j})},$$

где n – длина хромосомы. Таким образом, ожидаемое число повторов троек оснований (ijk) и $(\overline{k}\overline{j}\overline{i})$ совпадает по длине хромосомы. Симметрия для последовательностей оснований также подтверждается для модели однородной цепи Маркова и вытекает из симметрии пар оснований. Этот результат является следствием важного утверждения.

Утверждение 4. Оценка вероятности последовательности $x_1, x_2, ..., x_{n-1}, x_n$ совпадает с оценкой вероятности последовательности $\overline{x}_n, \overline{x}_{n-1}, ..., \overline{x}_2, \overline{x}_1$, т.е.

$$\hat{p}(x_1, x_2, ..., x_{n-1}, x_n) = \hat{p}(\overline{x_n}, \overline{x_{n-1}}, ..., \overline{x_2}, \overline{x_1}) \quad (12)$$

Отсюда следует, что вероятности двух противоположных нитей хромосомы, подсчитанные в модели однородной цепи Маркова на основе оценок переходных вероятностей (6), совпадают.

2. Одинаковая полярность цепей ДНК

Заметим, что симметрия оснований (1), (2) $n(i,1) = n(i,2)$, $i \in \{\mathrm{A,C,G,T}\}$ может выполняться и в том случае, когда обе комплементарные нити ДНК имеют одинаковые направления записи оснований (рис.2). Однако в природе такой вид симметрии отсутствует.

$$
\begin{array}{l}
5' \,\_\_\_\_\_\_\_\_ ij \,\_\_\_\_\_\_\_\_ \overline{i}\overline{j} \,\_\_\_ \rightarrow \_\_ 3' \quad \text{1-st string} \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \updownarrow\updownarrow \ldots\ldots\ldots\ldots\ldots \\
5' \,\_\_\_\_\_\_\_\_ \overline{i}\overline{j} \,\_\_\_\_\_\_\_\_ ij \,\_\_\_ \rightarrow \_\_ 3' \quad \text{2-nd string}
\end{array}
$$

Рис 2. Одинаковая полярность цепей ДНК

Поэтому из симметрии оснований (1), (2) нельзя вывести симметрию пар оснований. В таком случае симметрия пар оснований $n(ij,1) = n(ij,2)$, $i, j \in \{A,C,G,T\}$, вытекает из следующих соотношений, выполняющихся для одной нити:

$$n(AA)= n(TT),\ n(CC)= n(GG),\ n(AC)= n(TG),\ n(CA) = n(GT)$$

$$n(AG)= n(TC),\ n(CG) = n(GC),\ n(AT)= n(TA),\ n(CT) = n(GA) \qquad (13)$$

или в виде одной формулы

$$n(ij) = n(\overline{i}\ \overline{j}) \qquad (14)$$

Заметим, что в отличие от рассмотренной симметрии (4) пары AT, TA , CG, GC, присутствуют в (14), т.е. для пар оснований на два ограничения больше.

Для симметрии (14) ограничения (8), (9) трансформируются в одно ограничение

$$n(CA) + n(GA) = n(AC) + n(AG) \qquad (15)$$

Симметрия троек оснований $n(ijk,1) = n(ijk,2)$ вытекает из соотношений, выполняющихся для одной нити ДНК:

$$n(ijk) = n(\overline{i}\overline{j}\overline{k}) \qquad (16)$$

Для симметрии вида (16) добавляется два связывающих ограничения для троек оснований.

Для симметрии с одинаковой полярностью нитей ДНК справедливы рассмотренные выше утверждения 1–3. Из соотношений (1), (14) вытекает, что для однородной цепи Маркова оценки вероятностей троек оснований (ijk) и $(\overline{i}\overline{j}\overline{k})$ совпадают:

$$n\hat{p}(ijk) = \frac{n(i)n(ij)n(jk)}{n(i)n(j)} = n\hat{p}(\overline{i}\overline{j}\overline{k}) = \frac{n(\overline{i})n(\overline{i}\overline{j})n\left(\overline{j}\overline{k}\right)}{n(\overline{i})n(\overline{j})}.$$

Утверждение 4 записывается следующим образом.

Утверждение $4'$. Оценка вероятности последовательности $x_1, x_2, ..., x_{n-1}, x_n$ совпадает с оценкой вероятности последовательности $\overline{x}_1, \overline{x}_2, ..., \overline{x}_{n-1}, \overline{x}_n$, т.е.

$$\hat{p}(x_1, x_2, ..., x_{n-1}, x_n) = \hat{p}(\overline{x_1}, \overline{x_2}, ..., \overline{x_{n-1}}, \overline{x_n}).$$

Проведенные исследования показывают, что симметрия вида $n(ij) = n(\overline{j}\,\overline{i})$ имеет восемь связывающих ограничений (3), (8), (9) для пар оснований, а симметрия $n(ij) = n(\overline{i}\ \overline{j})$ содержит девять ограничений (13), (15). Для троек оснований у симметрии вида $n(ijk) = n(\overline{i}\overline{j}\overline{k})$ на два ограничения больше, чем у симметрии $n(ijk) = n(\overline{k}\,\overline{j}\,\overline{i})$. Поэтому ДНК с противоположной полярностью нитей имеет больше степеней свободы, чем ДНК с одинаковой полярностью, т.е. с точки зрения теории информации модель Уотсона-Крика более эффективна.

3. Генерация случайных последовательностей с симметриями обеих видов

С помощью модели цепей Маркова можно легко сгенерировать случайные последовательности, для которых будет выполняться симметрия вида (1), (4), (10) для модели Уотсона-Крика. На основе оценок переходных вероятностей (6) и программы псевдослучайных чисел строились случайные последовательности оснований различной длины, в том числе и совпадающие по длине с хромосомами человека. Аналогичным образом генерировались случайные последовательности, для которых выполняются симметрии вида (1), (14), (16) для модели с одинаковой полярностью цепей ДНК. Численные расчеты показали, что относительная разность между тройками оснований в (10) и (16) значительно меньше 1%. Таким образом, на основе модели цепей Маркова генерируются случайные последовательности, для которых выполняется симметрия для двух видов полярности нитей ДНК.

Заключение

Показано, что для нитей ДНК возможно два вида симметрии, но в природе реализована более эффективная с точки зрения теории информации модель Уотсона-Крика. Симметрия отдельных оснований – следствие симметрии пар оснований и соответственно симметрия пар оснований – следствие симметрии троек оснований. С помощью модели однородной цепи Маркова подтверждается, что симметрия последовательностей оснований вытекает из симметрии коротких последовательностей (пар оснований).

Решение сложных задач предсказания пространственной структуры белков показало, что если соотношения симметрии в записи генетической информации не выполняются, то байесовские процедуры распознавания на цепях Маркова не работают [2]. Полученные результаты открывают широкие возможности применения байесовских процедур на моделях цепей Маркова для распознавания свойств участков оснований (генов), в том числе генетических заболеваний.

Библиография

1. Гупал А.М., Вагис А.А. Комплементарность оснований в хромосомах ДНК // Проблемы управления и информатики. – 2005. – № 5. – С. 90–94.
2. Гупал А.М., Сергиенко И.В. Оптимальные процедуры распознавания.– Киев: Наукова думка, 2008.– 232 с.
3. Сергиенко И.В., Гупал А.М., Вагис А.А. Правила симметрии в записи генетической информации ДНК // Кибернетика и системный анализ. – 2011. – № 3. – С. 88–94.

Сведения об авторах

***Вагис Александра Анатольевна** – Институт кибернетики им. В.М. Глушкова НАН Украины, к.ф.-м.н., с.н.с., Киев, Украина. E-mail: valex_ic@mail.ru*

***Гупал Анатолий Михайлович** – Институт кибернетики им. В.М. Глушкова НАН Украины, д.ф.-м.н., чл.-к. НАН Украины, зав. отд., Киев, Украина. E-mail: gupal_anatol@mail.ru*

ПРОБЛЕМЫ СОЗДАНИЯ ЖИЗНЕСПОСОБНЫХ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ И МЕТОДЫ ИХ РЕШЕНИЯ

Валерия Грибова, Александр Клещев

Abstract: В данной работе рассматриваются новые идеи, связанные с созданием жизнеспособных интеллектуальных систем. Дана классификация задач управления, обеспечивающих их жизнеспособность, для каждого класса задач выделены проблемы, возможные методы их решения. Приведены результаты, полученные к настоящему времени по управлению интеллектуальными системами.

Keywords: интеллектуальные системы, онтология, база знаний, решатель задач, пользовательский интерфейс.

ACM Classification Keywords: D.1 Техника программирования D.2.2 Методы и средства проектирования D.2.10 Проектирование D.2.11 Программные архитектуры H.1 Модели и правила H.1.1 Теория систем и информации.

Введение

Исследования в области искусственного интеллекта продолжаются более сорока лет и к настоящему времени, с одной стороны, достигнуты значительные успехи, с другой – не наблюдается широкого и повсеместного использования интеллектуальных систем. Основные причины можно видеть в том, что они не соответствуют современным требованиям к ним: системы объяснения и ввода данных жестко встроены в код, их гибкая настройка на потребности пользователей не предусмотрена, базы знаний представлены в виде, непонятном экспертам предметной области, они не могут их сопровождать, в результате чего они быстро устаревают, а доверие к ним специалистов является очень низким. Эти причины ведут к тому, что большинство потенциальных пользователей предпочитают не использовать интеллектуальные системы.

Статистика утверждает, что за время жизненного цикла примерно треть всех усилий затрачивается на разработку программной системы, а две трети – на сопровождение, поскольку требования к ней различных пользователей и условия ее эксплуатации изменяются, что требует постоянной модификации системы. Для снижения трудоемкости модификации были введены понятия структурного программирования, модульности, абстракции данных, объектно-ориентированного программирования, а также сформулированы требования к «хорошему» структурированию кода (влияющему на сопровождаемость). Однако сопровождение программной системы по-прежнему остается очень трудоемким процессом. Учесть все возможные изменения требований и условий эксплуатации на стадии проектирования оказывается невозможным, поэтому после длительных изменений старого программного средства, как правило, возникает необходимость разработки нового, которое сталкивается с теми же проблемами. Чтобы избежать ситуации, когда затраты на сопровождение программной системы могут превысить выгоды от ее использования, программная система должна быть управляемой или жизнеспособной (адаптируемой и адаптирующейся). Для этого было предложено сопровождение, под которым понимается изменение ее кода, заменить управлением, под которым понимается решение задач сопровождения программной системы с помощью специальных высокоуровневых механизмов управления, сводящих к минимуму изменение ее кода [Norvig, 1998]. В основе технологии разработки таких систем лежит принцип: все проектные решения, влияющие на пользовательские свойства системы, должны совершенствоваться в процессе накопления опыта ее использования. Эти идеи вызвали оживленную дискуссию среди программистов, однако не нашли никакого отклика в литературе по интеллектуальным системам, хотя именно там их применение может быть особенно эффективным.

Целью работы является обсуждение проблем жизнеспособности интеллектуальных систем и возможных методов их решения.

Концепция обеспечения жизнеспособности интеллектуальных систем

Традиционные программные средства, детерминировано формирующие свой выход по входу, обычно рассматриваются как черный ящик. Адаптируемые программные средства, допускающие управляющие воздействия, обычно называют открытыми, а имеющие в своем составе и средства управления с обратной связью, формирующие управляющие воздействия, – адаптирующимися или жизнеспособными программными средствами (см. рис. 1)

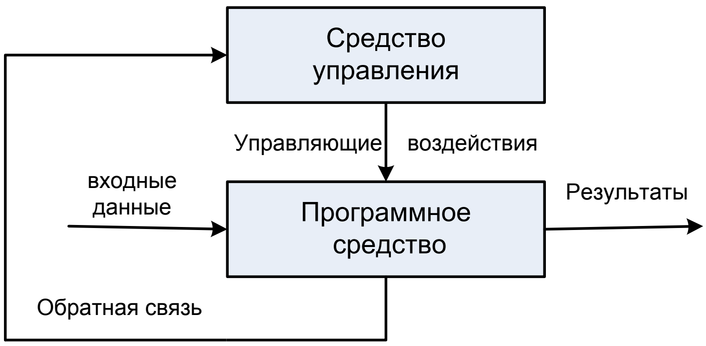


Рис. 1. Система управления программным средством

В основе предлагаемой концепции жизнеспособности интеллектуальных систем лежит принцип управляемости, то есть предлагается заменить сопровождение интеллектуальной системы управлением ею в процессе жизненного цикла [Грибова В.В.и др., 2010].

Архитектуру интеллектуальной системы в первом приближении можно представлять себе как взаимосвязанную тройку, состоящую *из базы знаний (и других информационных ресурсов – баз данных, онтологий, метаонтологий и т.п.), решателя задач*, реализующего функциональность системы, и *пользовательского интерфейса*. Цели управления этими тремя компонентами интеллектуальной системы в общем случае различны и в определенной степени независимы. В соответствии с этим в архитектуре системы, управляющей интеллектуальной системой, будем выделять три подсистемы: подсистему управления информационными ресурсами, подсистему управления решателем задач и подсистему управления пользовательским интерфейсом.

Рис. 2 Классификация задач управления интеллектуальными системами

Исходя из того, что в архитектуре интеллектуальной системы, как объекта управления, можно выделить три самостоятельных компоненты (информационные ресурсы, в том числе базу знаний, решатель задач и пользовательский интерфейс), а в системе, управляющей программной системой, три типа подсистем: ручного управления (интерактивное изменение свойств программной системы человеком), автоматического управления (без участия человека изменение свойств программной системы) и автоматизированного управления (смешанный тип управления), задачи управления интеллектуальными системами можно разделить на 9 классов (рис. 2).

Ручное (интерактивное) управление интеллектуальными системами

Ручное управление компонентами интеллектуальной системы направлено на приведение их содержания в соответствие с представлениями об этом содержании у лиц, управляющих этими компонентами.

Можно выделить три основные проблемы, связанные с ручным управлением интеллектуальными системами:

1. **Разнородные классы объектов управления: информационные ресурсы, решатели задач, пользовательские интерфейсы.** Задача интерактивного управления базами знаний была решена отечественными и зарубежными учеными еще в конце 80-х-начале 90-х годов прошлого века. Однако эти механизмы не были обобщены на случай интерактивного управления другими классами информационных ресурсов (например, онтологиями, метаонтологиями), а механизмы интерактивного управления программными компонентами (решателями задач и пользовательским интерфейсом) в литературе вообще не рассматривались.

2. **Разнородный состав группы управляющих для каждого класса объекта управления.** Различными компонентами интеллектуальных систем должны управлять специалисты, относящиеся к различным предметным областям: базами знаний и данных – эксперты и специалисты соответствующих предметных областей, онтологиями – инженеры знаний, программными компонентами – программисты, интерфейсными компонентами – дизайнеры, эргономисты и т.п. Механизмы интерактивного управления, понятные всем этим различным группам управляющих, в литературе также не обсуждались.

3. **Достижение согласованности управления интеллектуальными системами.** Компоненты интеллектуальной системы связаны между собой и независимое управление этими компонентами может нарушить связи межу ними, поэтому при управлении интеллектуальной системой необходимо обеспечить согласованность между различными ее компонентами.

Для решения указанных выше проблем *предлагаются следующие решения:*

- представлять все компоненты интеллектуальной системы в виде декларативных информационных ресурсов;

- разработать общий подход к интерактивному управлению информационными ресурсами, при котором с каждым информационным ресурсом связывается своя формально представленная концептуальная система;

- представлять связи между компонентами интеллектуальных систем также в виде декларативных информационных ресурсов.

В информационном ресурсе важна информация и ее удобное представление для управляющего, а не способ компьютерного представления, т.е. рассматриваются только концептуальные информационные ресурсы, которые могут обрабатываться интеллектуальными системами, а структура и содержание которых понятны управляющим ими. Из вышесказанного следует, что концептуальные информационные ресурсы, в том числе и базы знаний, имеют декларативное и структурное представление (такой информационный ресурс можно представлять в виде семантической сети, структура которой определяется некоторой метаинформацией, а каждый термин, вводимый в этой сети, относится к некоторому классу, определяемому этой метаинформацией).

Метаинформацию концептуального информационного ресурса можно рассматривать как его онтологию, как грамматику, порождающую этот ресурс, как сценарий диалога с пользователем при редактировании этого ресурса и как основу языка запросов к этому ресурсу со стороны интеллектуальной системы. Концептуальный информационный ресурс является неразрывной парой, состоящей из метаинформации и собственно содержания (контента) этого ресурса, которые находятся между собой в определенном соответствии. Представление баз знаний и других информационных ресурсов в виде семантических сетей в настоящее время является почти общепринятым. На рис. 3 приведен фрагмент медицинской базы знаний в форме семантической сети.

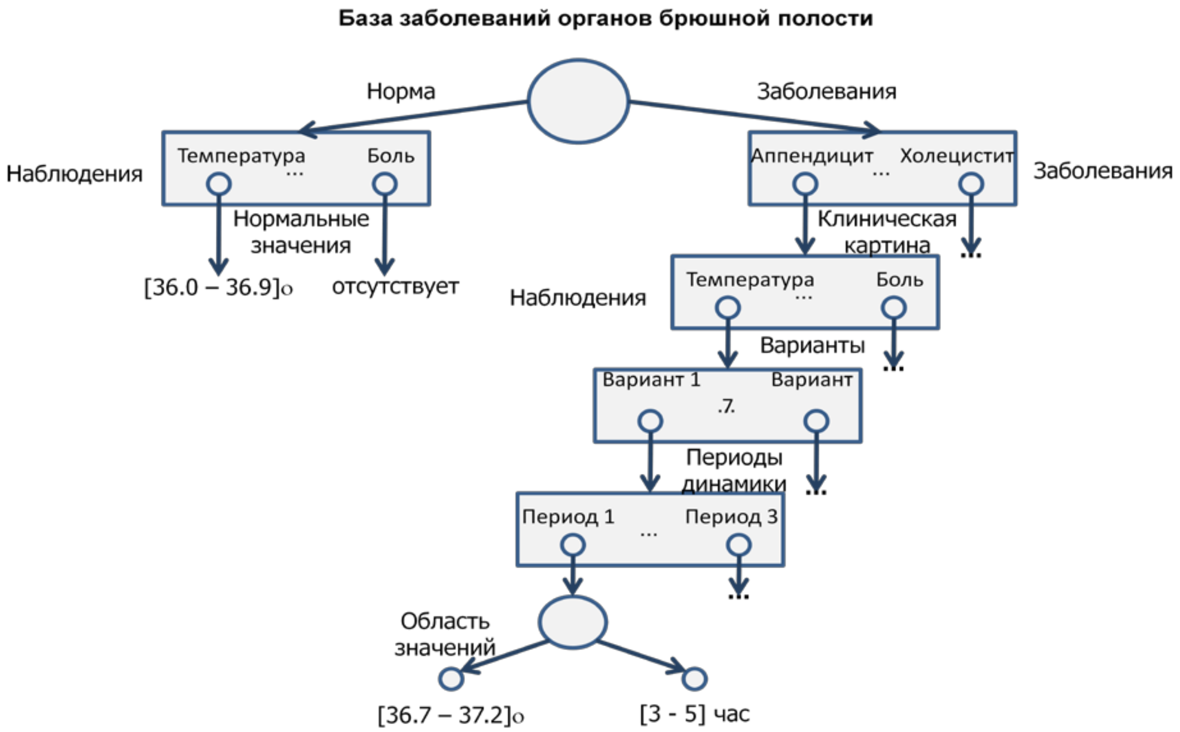


Рис. 3 Представление базы знаний в виде декларативного информационного ресурса (семантической сети)

В настоящее время все большую популярность приобретает технология, при которой решатель задач интеллектуальных систем строится из агентов. В работе [Клещев А.С., 2009] введено представление агентов в виде декларативно-процедурных информационных ресурсов: решатель состоит из множества блоков, а каждый блок состоит из онтологии сообщений и множества продукций. При этом каждый решатель состоит из трех информационных ресурсов:

– сети агентов, определяющей, какие агенты входят в решатель,

- управляющего графа, позволяющего агентам находить адресатов посылаемых ими сообщений,

- схемы распределения, устанавливающей ограничения на распараллеливание агентов при исполнении решателя в многопроцессорной среде (может отсутствовать) [Крылов Д.А., 2010].

В работах [Грибова В.В., 2005], [Грибова В.В., 2006] показано, что модель пользовательского интерфейса также может быть представлена в виде декларативного информационного ресурса, состоящего из следующих основных компонентов:

- модели задач,
- модели предметной области,
- модели пользователя,
- модели представления
- модели сценария диалога.

По декларативно представленной модели универсальный процесс интерпретации полностью определяет поведение интерфейса [Грибова В.В., 2008]. Таким образом, все компоненты интеллектуальной системы и связи между ними могут быть представлены в виде семантических сетей – декларативных информационных ресурсов.

Концептуальная система информационного ресурса входит в его метаинформацию и позволяет представить информационный ресурс в форме, понятной для управляющего им, через процесс визуализации (в форме семантической сети или текста). На рис. 4 приведено формальное представление (на структурно-логическом языке) концептуальной системы медицинской базы знаний. Таким образом, для включения в процесс управления специалистов различных предметных областей, с каждым информационным ресурсом связывается своя формально представленная концептуальная система.

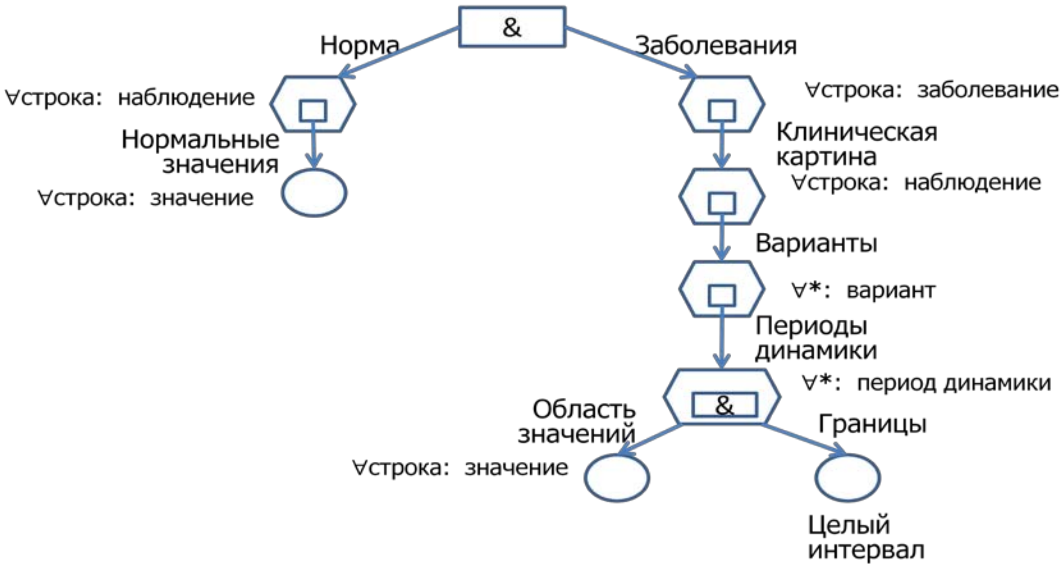


Рис. 4 . Пример концептуальной системы декларативного информационного ресурса

Средством управления (созданием и редактированием) декларативными информационными ресурсами (онтологиями, базами знаний и данных) является разработанный универсальный редактор классов семантических сетей, который интерпретирует метаинформацию информационного ресурса (метаинформацию создает инженер знаний) и генерирует интерфейс для управляющего этим информационным ресурсом [Клещев А.С., 2006].

Автоматическое управление интеллектуальными системами

Проблемами, связанными с автоматическим управлением интеллектуальными системами, являются:

1. Разнородные классы объектов управления: информационные ресурсы, решатели задач, пользовательские интерфейсы. Данная проблема является общей для всех задач управления, метод ее решения рассмотрен выше.

2. Разнородность состояний объекта управления (автоматическое управление в периоды функционирования объекта и между этими периодами). Данная проблема состоит в том, что наряду с задачами ручного и автоматического управления интеллектуальной системой в периоды, когда она не функционирует, возникают задачи автоматического управления ею во время ее функционирования.

3. Достижение согласованности механизмов ручного управления с механизмами автоматического управления. Механизмы автоматического управления должны быть согласованы с механизмами ручного управления, чтобы управляющие могли контролировать процессы автоматического управления.

4. Различная степень неопределенности в задачах автоматического управления.

Разнородность классов объектов управления пока не позволила найти общих, проблемно-независимых решений для задач автоматического управления, поэтому в настоящее время предлагается использовать и проблемно-зависимые решения.

Так, для автоматического управления базами знаний к настоящему времени разработаны лишь проблемно-зависимые механизмы (для онтологии медицинской диагностики). Примером такого механизма автоматического управления с обратной связью является итеративный алгоритм индуктивного формирования медицинской базы знаний. Он состоит из двух асинхронных процессов (см. рис. 5) - первый из них обрабатывает поступающие один за другим примеры из обучающей выборки и после обработки некоторой порции новых примеров передает результат своей работы (множество альтернативных баз знаний) второму процессу для выбора из этого множества наилучшей базы знаний и обновления рабочей базы знаний [Клещев А.С., 2006].

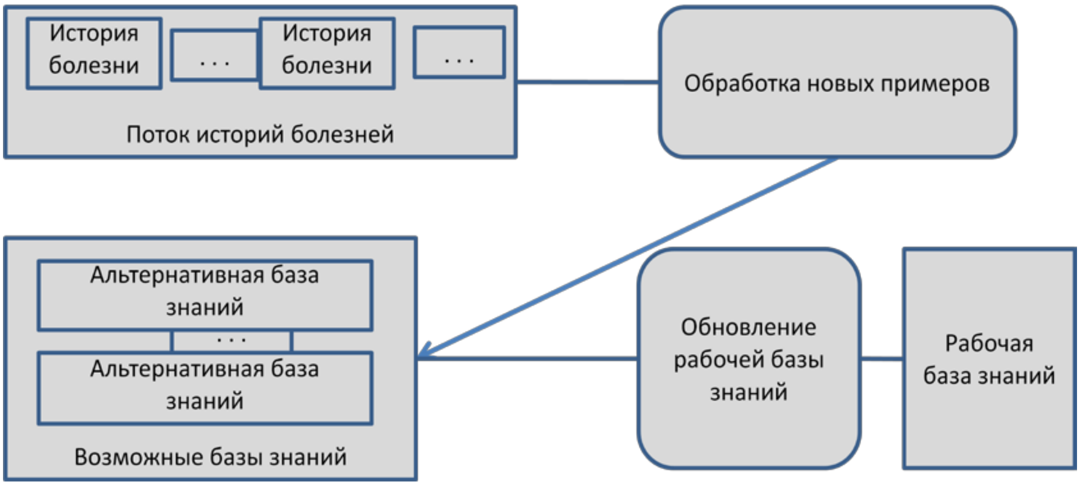


Рис. 5 Итеративный метод индуктивного формирования баз медицинских знаний

Для автоматического управления пользовательским интерфейсом разработаны проблемно-независимые механизмы – механизм автоматической адаптации к данным и особенностям пользователей [Грибова В.В., 2010]. Этот механизм на основе динамических характеристик пользователя, а также формируемых решателями задач наборов данных формирует визуальное представление диалога, соответствующее как требованиям юзабилити, так и индивидуальным особенностям пользователей системы. Для автоматического управления решателями задач предложены как проблемно-зависимые, так и проблемно-независимые механизмы.

Примером проблемно-независимого механизма в решателе задач является автоматическое распараллеливание функционирования агентов на многоагентной платформе. Управление осуществляется на четырех уровнях – интеллектуальные системы функционируют параллельно друг с другом, запускаемые через информационно-административную систему на основе информации о полномочиях пользователей; агенты внутри каждой интеллектуальной системы взаимодействуют друг с другом асинхронно с помощью коммуникационной системы на основе управляющего графа; распараллеливание агентов внутри интеллектуальной системы управляется схемой распределения агентов в интеллектуальной системе; а равномерная нагрузка на узлы вычислительной фермы (сервер, клиентские компьютеры, многопроцессорные системы) достигается алгоритмом балансировки [Крылов Д.А., 2010].

Примером проблемно-зависимого механизма является автоматическое управление решателем задач экспертной системы медицинской диагностики. На рис. 6 схематично представлен механизм управления таким решателем, когда результаты мониторинга базы знаний, выполняемого каждый раз после ее изменения, позволяют сократить множество проверяемых гипотез при обработке решателем каждой истории болезни на основе жесткой системы правил.

Рис. 6 Автоматическое управление решателем задач экспертной системы медицинской диагностики

Решением второй проблемы (два состояния объекта управления), является использование наряду с внешними механизмами также и внутренних механизмов автоматического управления. Внешние механизмы обеспечивают автоматическое управление базами знаний между периодами функционирования интеллектуальных систем (примером является рассмотренный выше итеративный метод индуктивного формирования баз медицинских знаний), а внутренние механизмы – автоматическое управление решателями задач (механизм автоматического распараллеливания функционированием агентов на многоагентной платформе) и пользовательскими интерфейсами (механизм автоматической адаптации к данным и особенностям пользователей) в периоды их функционирования.

Проблема достижения согласованности механизмов автоматического управления с механизмами ручного управления осуществляется через использование метаинформации в механизмах внешнего управления. Ранее отмечалось, что использование метаинформации позволяет представлять информационные ресурсы в виде, понятном для управляющих. Использование той же метаинформации позволяет управляющим следить и за процессом автоматического управления информационными ресурсами, например, базами знаний.

Решением проблемы различной степени неопределенности в механизмах автоматического управления является использование систем правил и механизмов с обратной связью. Так, если степень неопределенности в задаче автоматического управления не высока, то механизм автоматического управления может строиться на основе системы правил, но в случае высокой степени неопределенности в некоторых задачах приходится разрабатывать механизмы автоматического управления с обратной связью. Среди рассмотренных выше примеров итеративные методы индуктивного формирования знаний и автоматической адаптации к особенностям пользователей построены на основе механизмов с обратной связью. С использованием систем правил реализованы механизмы автоматического распараллеливания функционированием агентов и автоматической адаптации к данным.

Автоматизированное управление

Проблемами, связанными с автоматизированным управлением интеллектуальными системами, являются:

1. Разнородные классы объектов управления: информационные ресурсы, решатели задач, пользовательские интерфейсы.

2. Разнородный состав группы управляющих для каждого класса объекта управления.

3. Объединение механизмов ручного и автоматического управления. Данная проблема состоит в том, как организовать взаимодействие между управляющим и механизмами автоматического управления.

Решение первых двух проблем возможно через совместное использование механизмов, разработанных для ручного и автоматического управления. Объединение же механизмов ручного и автоматического управления либо дает возможность управляющему корректировать результаты автоматического управления, либо через интерактивные механизмы показывать решение частных задач, которые затем обобщаются механизмами автоматического управления и используются решателем.

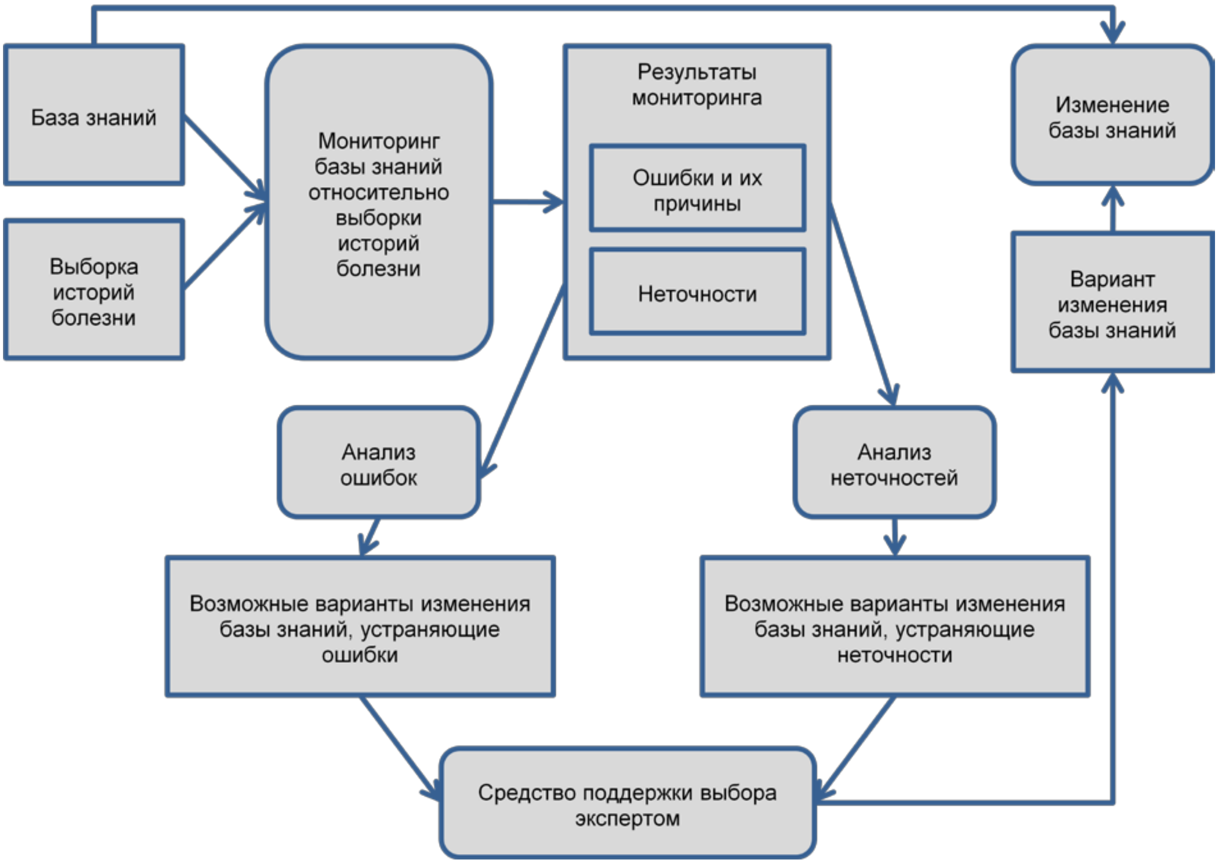


Рис. 7 Схема отладки базы медицинских знаний.

Примером задачи автоматизированного управления может служить задача отладки баз медицинских знаний, где мониторинг базы знаний относительно выборки историй болезни выявляет ошибки и неточности в этой базе знаний, механизмы индуктивного формирования баз знаний определяют возможные варианты устранения этих ошибок и неточностей, а управляющий базой знаний эксперт выбирает из этих вариантов те, которые он считает наилучшими. (см. рис. 7).

Заключение

Жизнеспособность интеллектуальных систем предполагает, что она должна содержать не только базу знаний, метод решения задач и способы взаимодействия с пользователями, но также и модели механизмов управления ими. При этом коллектив управляющих интеллектуальной системой, состоящий из экспертов предметной области, инженеров знаний, дизайнеров и программистов должен целенаправленно улучшать свойства интеллектуальной системы и адаптировать ее к изменениям, происходящим в процессе жизненного цикла. При этом сам процесс управления интеллектуальными системами, формирование команды управляющих и правила ее работы являются новыми и требуют экспериментального изучения. Необходимо там, где это возможно, обобщать опыт управления интеллектуальными системами и разрабатывать проблемно-независимые модели, механизмы и средства управления, а также заменять ручное управление на автоматическое и автоматизированное. В настоящее время в коллективе, в котором работают авторы, разработаны и используются некоторые как проблемно-зависимые, так и проблемно независимые механизмы для ручного, автоматического и автоматизированного типов управления, работы в данном направлении продолжаются.

Благодарности

Работа выполнена при финансовой поддержке РФФИ, проект "Управление концептуальными метаонтологиями, онтологиями, знаниями и данными в интеллектуальных системах" (10-07-00089-а) и ДВО РАН в рамках Программы №2 Президиума РАН, проект "Развитие систем управления базами знаний с коллективным доступом" (09-I-П2-04)

Литература

[Norvig, 1998] P. Norvig, D. Cohn. Adaptive software.URL: http://norvig.com/adapaper-pcai.html

[Грибова В.В.и др., 2010] Грибова В.В., Клещев А.С., Шалфеева Е.А. Управление интеллектуальными системами // Известия РАН. Теории и системы управления.- 2010.- № 6.- С. 122-137.

[Клещев А.С., 2009] Клещев А.С. Концепция многоагентной системы в многоцелевом компьютерном банке знаний // Четвертая международная конференция по проблемам управления: Сборник трудов. М.: Учреждение РАН Институт проблем управления имени В.А.Трапезникова РАН.- 2009.- С. 1585-1595. ISBN 978-5-91450-026-6 [Электрон.].

[Крылов Д.А., 2010] Крылов Д.А. Облачная платформа для создания и управления интеллектуальными Интернет-сервисами // Инфокомуникационные и вычислительные технологии и системы: материалы III Междунар. конф. Улан-Удэ: Изд-во Бурятского госуниверситета.- 2010.- С. 180-183. ISBN 978-5-9793-0282-9

[Грибова В.В., 2005] Грибова В.В., Клещев А.С. Использование методов искусственного интеллекта для проектирования пользовательского интерфейса // Информационные технологии.- 2005.- №8. - С.58-62.

[Грибова В.В., 2006] Грибова В.В., Клещев А.С. Управление проектированием и реализацией пользовательского интерфейса на основе онтологий // Проблемы управления.- 2006.- №2.- С.58-62.

[Грибова В.В., 2008] Грибова В.В. Модель генерации кода пользовательского интерфейса для различных типов диалога // Научно-технические ведомости СПбГПУ. – Санкт-Петербург: Изд-во Политехнического университета. – 2008. – №3. – С.145-151. – ISSN 1994-2354.

[Клещев А.С., 2006] Клещев А.С., Орлов В.А. Компьютерные банки знаний. Универсальный подход к решению проблемы редактирования информации// Информационные технологии.- № 5.- 2006.- С. 25-31.

[Клещев А.С., 2006] Клещев А.С. Задачи индуктивного формирования знаний в терминах непримитивных онтологий предметных областей. Научно–техническая информация.-Серия 2.- 2003.- № 8- С. 8-18.

[Грибова В.В., 2010] Грибова В.В., Черкезишвили Н.Н. Развитие онтологического подхода для автоматизации разработки пользовательских интерфейсов с динамическими данными // Информационные технологии.- 2010.- №10.- С. 54-58.

Информация об авторах

Валерия Грибова — д.т.н., зав. лабораторией интеллектуальных систем Института автоматики и процессов управления Дальневосточного отделения Российской академии наук, г. Владивосток, ул. Радио, 5, тел. +7 (4323) 314001, gribova@iacp.dvo.ru, http://www.iacp.dvo.ru/is.

Александр Клещев — д.ф.-м.н., главный научный сотрудник отдела Интеллектуальных систем Института автоматики и процессов управления Дальневосточного отделения Российской академии наук, г. Владивосток, ул. Радио, 5, тел. +7 (4323) 310421, kleschev@iacp.dvo.ru, http://www.iacp.dvo.ru/is.

ФИЗИКО-ОНТОЛОГИЧЕСКИЙ ПОДХОД К ПОСТРОЕНИЮ ЦЕЛОСТНОЙ КАРТИНЫ МИРА

Мержвинский Анатолий Александрович

Abstract: На основе анализа физики взаимодействий материальных объектов предложена крупноблочная мета-модель мира в виде множества E-i-E-i-E компонент, содержащих каморки, в каждой из которых могут быть отображены объекты универсума. Каморки объединены в модули. 3-мерная композиция модулей отображает основные категории бытия и процесс взаимодействия объектов. Приведен пример конкретной модели, отображающей 16 категорий материальных и информационных объектов. Модель может быть реализована программно либо выполнена физически в виде кубуса, состоящего из 17 модулей.

Keywords: universum, OBJECT, INTERACTIONS, IMAGE, COMMUNICANT

ACM Classification Keywords: GIT 2011, H.1.1 Systems and Information Theory; I.2.4 Knowledge Representation Formalisms and Methods.

Введение

Способность разума охватить всю вселенную своим мысленным взором ограничена и разработка методов отображения предметных областей (ПдО) – актуальная задача, решаемая во многих науках. В качестве предела сложности выставляются модели сложных иерархических систем конкретных ПдО, а также целеустремленных систем. Функционирует универсальная Интернет-энциклопедия Википедия. Ставится задача создания базы знаний для роботов, которой смогут пользоваться роботы в любой точке земного шара. Задача создания целостного образа мира, которая ставилась еще древними философами, по-прежнему в состоянии обсуждения на различных форумах. Созданию «картин мира» и «целостного образа мира», посвящено много публикаций, однако даже определение этих понятий является открытым. В частности, это объясняется следующими факторами:

Неравномерное развитие различных областей знаний. Разработка теорий определенных ПдО и взрывной характер развития компьютерной техники хотя и привели к успешному созданию проблемно-ориентированных информационных систем (ИС), однако, на наш взгляд, увели исследователей от проблем создания и визуализации обобщенной модели *универсума*, как высшей ступени объединения объектов ПдО.

Существующие физические теории материального мира и логические теории представления знаний отделены друг от друга и, развиваясь относительно независимо, на наш взгляд, недостаточно отражают природу информационных объектов, их связь с материальными объектами.

- Преимущественно развивалась теория передачи, приема и обработки информации, а теория устройств ввода-вывода и процессы сбора, ввода, вывода, как направление теории «отражения» выпали из общей теории информации.
- Понятия n-мерного отражения действительности переносятся на свойства материального трехмерного мира.

Нарушение принципа «не плоди ненужного» привело к дублированию либо многозначности терминологии в одних областях и отсутствии ключевых терминов – в других, а в ряде случаев и к выпячиванию никчемных проблем.

Существующие подходы к отображению действительности и их недостатки

Формируемая в результате познавательной деятельности человека картина мира складывается из представлений о различных ПдО и характерна как для каждого индивидуума, так и создаваемых ИС. Представление ментальных структур, изоморфных связям и отношениям между фрагментами реального мира, в машинной форме – основная задача создания баз знаний [Кургаев , 2008].

Известны различные картины мира, основанные на введении характерных категорий объектов и отношений между ними: *физическая*, посвященная миру вещей и традиционно ограничиваемая рамками неживой природы, *философская*, отображающая мир идей, *социо-психологическая* отображающая мир людей [Боровиков, 2002] и множество других.

Известен *научный* подход: мир - это Вселенная во всей ее совокупности форм материи в земном и космическом пространстве, т.е. все то, что существует вокруг нас [Степин, 2004]. *Научная* картина мира - в широком смысле - теоретическое отражение различных сторон действительности. Основные используемые понятия: *объект* – неоднородность распределения энергии в пространстве; *структура* - отображает пространственный аспект; *функция* – временной; *онтология* – отображает бытие. Понятий структура и функция - недостаточно для описания цепочки взаимодействий. Последняя представляется в виде «*процесса» и* удобно отображается с помощью семантической сети.

Системный подход – ориентируется на выбор объекта, как некоторой целостности, его внутренние и внешние связи. По степени информированности исследователя об объекте существует деление объектов на три типа «ящиков»:

- «белый ящик»: об объекте известно все;
- «серый ящик»: известна структура объекта, неизвестны количественные значения параметров;
- «черный ящик»: об объекте неизвестно ничего.

Для описания систем используется ER-модель (сущность-отношение); однако системный подход не рассматривает процессы отображения объектов и их сетевых взаимодействий.

Онтологическая картина мира может быть представлена на основе онтологий верхнего уровня, ориентированных на традиционные предметные области. Онтологии описываются тройками, шестерками, в которых первым множеством идет множество понятий либо классов понятий [Палагин, 2009]. Однако универсум включает не только понятия, но и материальные объекты, поэтому в модель универсума наряду с отображениями свойств объектов логично внести конкретные материальные объекты или множества объектов, которые могут определяться цифро-буквенным идентификатором и именем [Лозовский, 2003], которые, в частности, могут включать качественные и количественные характеристики.

Попытки "увязать картину мира во что-то целое, что подвергается конструктивному анализу и пригодное для практического использования" были сделаны, выполнена декомпозиция ноосферы на «физикал», «ментал» и « культурал» [Лозовский, 2003], разработаны методы и алгоритмы интеграции онтологий различных ПдО [Палагин, 2009], распространены различные категориальные картины мира, однако мечта Лапласа: «если бы удалось связать в единое целое все знания о мире» не выполнена. Общий недостаток рассмотренных методов – отображение некоторого среза либо части действительности при отсутствии охвата целостной картины мира. Таким образом, крупноблочная модель мира верхнего уровня, которая объединяет и наглядно отображает основные категории бытия [Лозовский, 2003], до сих пор не разработана.

Известные подходы к решению задачи отображения ПдО и их недостатки

Применительно к рассматриваемой задаче используемое в онтологии и семантических сетях базовое понятие «отношение» – очень широкое. Отношениями могут описываться как непосредственные, так и опосредствованные взаимодействия между объектами. При этом высокий уровень абстракции базовых понятий «объект» и «отношение» приводит к увеличению количества степеней свободы и соответственно числа возможных вариантов структур. Однако в материальном мире взаимодействия ограничены физическими свойствами объектов и физическими законами. Отличительным свойством материальных объектов является способность обмениваться со средой *V* и другими материальными объектами веществом *G*, излучением *Г* разной природы, энергией *E* и импульсом [Рис. 1]. Интегральное воздействие может быть измерено силой F_f либо импульсом силы $S_f = F_f * t$. Отход от чрезмерного абстрагирования и введение более узких физических понятий, отражающих физические законы взаимодействия, сокращает число степеней свободы и, таким образом, потенциально упрощает видение возможных вариантов реализации процессов. Такой подход к отображению реальности позволяет определить его как физический.

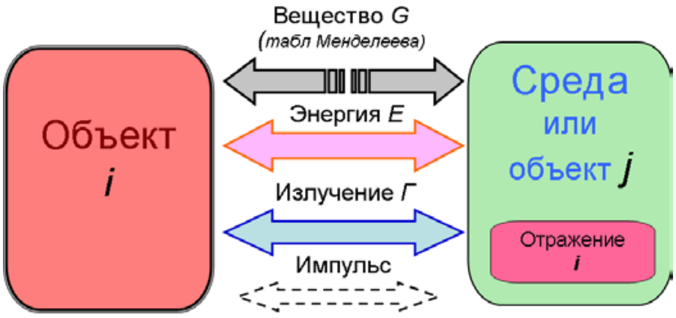


Рис.1. Важнейшие свойства материальных объектов

Детальность отображения объектов и связей объектов ПдО увеличивается по мере использования все более совершенных методов отображения: в виде гипертекста, коллекции глоссариев, онтологии, математических соотношений. Верхней ступенью, на наш взгляд, является представление ПдО и, естественно, универсума, в виде моделей, выражаемых всеми доступными средствами, в частности, с помощью предлагаемых ментально- и машинно-ориентированных *физико-онтологических моделей.* Взаимосвязь указанных моделей *U* с отображаемой реальностью представлена на рис. 2, на котором 1 – универсум *U* (совокупность объектов и явлений в целом, рассматриваемая в качестве единой системы); 2 –эксперт, имеющий некоторое ментальное представление об универсуме (как ПдО) в виде знаний, которые могут быть отображены в ментально-ориентированной модели 3. Переход от предметно-образной модели левого полушария к понятийно-образной модели правого полушарий реализуется с помощью некоторого научного сообщества 4 и сенсорных устройств 8, например, в виде устройств ввода/вывода (УВВ). Модель 3 может существовать в виде метамодели либо быть наполненной некоторым содержанием ПдО или *U* в виде конкретной формальной либо неформальной модели.

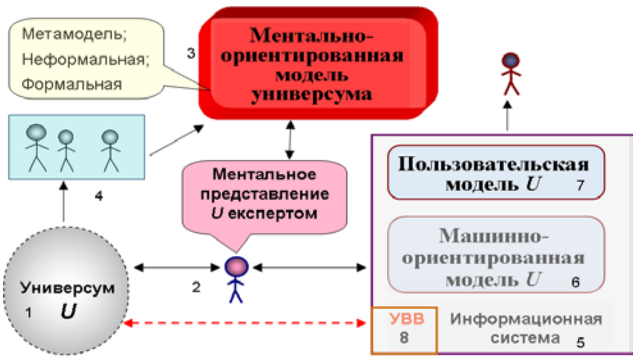


Рис. 2.. Взаимосвязь моделей универсума *U*

Предполагается, что ментально-ориентированная модель, включающая образы объектов ПдО, описания понятий, математические тексты, описывающие динамику взаимодействий, и пр. может быть введена экспертом в информационную систему 5 в форме машинно-ориентированной модели 6, в виде универсальных моделей представления знаний, применимых для большинства проблемных областей (семантических сетей, фреймов, продукционных систем и логических моделей) либо ориентированной на пользователя модели 7. В конкретных моделях также могут присутствовать отображения объектов универсума, введенные с помощью УВВ 8.

При разработке картины мира «сверху» естественным является использование таких наиболее абстрактных понятий как «Универсум» и «Модель» ибо «они, как всякая абстракция высшего уровня является результатом раскрытия все более существенных свойств вещей и явлений через их связи и отношения». Картина же мира – предполагает некоторую полноту. Модель может отображать некоторую сторону действительности. Вместо картины мира предлагается рассматривать *модель мира*, а при наличии всех категорий объектов - модель *универсума*.

Совершенствование технологий микро- нано- оптоэлектроники и голографии позволяет по-новому подойти к визуализации данных и знаний о трехмерных, а в общем случае и n-мерных сущностях объектов на основе цветных трехмерных отображений. С учетом изложенного цель настоящей работы: дать формальную модель мира, которая:

состояла бы из ограниченного небольшого числа понятий и не очень больших объяснений к ним;

согласовывалась с различными мировоззрениями (системой взглядов на природу и общество);

- позволяла бы путем ее наполнения построить интегральную «картина мира» в соответствии с мировоззрением ее творца и допускала бы постепенное расширение целостного образа мира.

Исходные положения о свойствах макромира

Вещество — вид материи, обладающей массой покоя. *Поле* — основной вид материи, связывающий частицы и тела. *Объекты:* в самом общем виде - неоднородности распределения материи и энергии в пространстве и времени. *Материальный* объект - выделенная для рассмотрения часть материального мира.

В реальной действительности, пространственной формой обладают только реальные тела. Виртуальные, нереальные тела, хоть и облечены в пространственную форму, протяженностью не обладают. Четырех - и более - мерных реальных образований в реальном пространстве также еще никто не обнаружил. Они существуют только в фантазиях и измышлениях [*]. Поэтому будем исходить из следующего:

1) Объекты материального макромира взаимодействуют в трехмерном пространстве; эти взаимодействия могут быть отображены в n-мерных моделях;

2) n-мерными ментальными моделями могут быть:

- Синтезированные сознанием модели (n-мерный куб, n-мерные таблицы, …)
- Отображения объектов материального мира.

Если более кратко: мир трехмерный; отображения мира могут быть многомерными.

Классы и ипостаси материальных объектов. Применительно к происхождению множество материальных объектов R формально опишем тройкой $R=<F, W, \Omega>$, где F - класс материальных неживых естественных объектов; W - класс материальных неживых искусственных объектов; Ω - класс материальных живых объектов.

В данной работе представляет интерес рассмотрение материальных объектов в ипостаси носителей «отражений» других объектов или «информации» о других объектах. Однако существует традиционное определение «Информацио́нный объе́кт» — информация, не зависящая от носителя и развивающаяся по собственным законам, пребывающая в информационном пространстве [Переслегин , 2001]. Поэтому в данной работе материальный объект в комплексе с содержащейся в нем информацией независимо от формы представления определим как *материальный информационный объект* (i_{mat}). В зависимости от вида операции взаимодействия с другими объектами I_{mat} могут:

- отображать некоторую сущность (идею, идеальный объект) других объектов (прообразов) *U*
- представлять самостоятельную сущность (идею, идеальный объект), которая может непосредственно или через объекты некоторой конфигурации Ф влиять на третьи объекты Q.

В первом случае, как отмечалось выше, это может быть "образ" объекта, а в более сложном, "модель" объекта. В случае же самостоятельной сущности это "код", "программа" или "модель".

Формально множество материальных информационных объектов *I* представим тройкой $I_{mat} = <X,\Theta,\Psi>$, *X* - класс элементов x, в котором элементы x - образы сущностей материального мира *R*, полученные с помощью инструментов (фото, показание измерительных приборов и др.), Θ -класс, в котором θ - продукты умственной деятельности, которые зафиксированы на любом носителе, Ψ - класс, в котором ψ - ментальные объекты в живом организме. Примеры характерных актантов (действующих участников взаимодействий) приведены в табл. 1

Таблица 1. Характерные примеры актантов

	Коммуникатор	Коммуникат	Свойства носителя	Оперант	Реципиент
	Тракт «материальный объект» - «сенсор»				
1	Материальный объект	Скалярный поток фотонов	Интенсивность P(λ)	Световод	Анализатор спектра
		Векторный поток фотонов	Сила излучения I(Ω)	Объектив	Фотоприемная матрица
Тракт «материальный объект» - «реципиент»					
2	Материальный культуральный объект	Поток бумажных носителей	Способность быть считанным	Транспортное средство (например, железная дорога)	Адресат
Тракт «Человек – Информационный объект»					
3	Идеальный объект	Силовое воздействие	Сила F(x,y) Возможность выбора x,y	«Оператор-клавиатура» (джойстик),	Память
		Звуки S(t)	Спектр S(f,t)	Микрофон	Память
Канал связи информационных объектов					
4	Информационный объект	Скалярный поток носителей	Мощность потока P(t)	Линия связи	Пороговое устройство (АЦП)
		Векторный поток носителей	Сила излучения	Видеоканал	Оператор
Носители ячеек памяти					
5	Информационный объект	Материальные кластеры	Электрическая проводимость	Схема програм-мирования ПЗУ	Матрица ПЗУ
			Коэффициент отражения	«Драйвер- лазер записи»	Оптический диск
		Электромагнитные домены	Магнитный поток	«драйвер-головка записи»	Магнитный диск
		Электрический ток	Величина тока ячейки	Цепь включения триггера	Схемная ячейка памяти

В зависимости от рассматриваемой ипостаси коммуникаторов и реципиентов - как материальных *R* или как информационных *I* (носителей или идеальных) - в акте взаимодействия могут принимать участие такие комбинации *коммуникантов: R-R, R-I, I-R, I-I.* Эти комбинации определяют категории *актов* взаимодействий коммуникантов (для выполнения требуется время), *операций* (последовательности актов опосредствованных взаимодействий с обратном связью) или *процессов* (последовательности операций взаимодействий). В случае опосредствованных актов взаимодействий результат может зависеть от влияния на оперант φ управляющих, координирующих или иным образом влияющих объектов (объектов-инфлюантов). Объекты-инфлюанты могут включат среды, временные синхронизаторы, программы и модели желательных и реальных процессов. Таким образом, процесс может рассматриваться как объект, который существует и в пространстве и во времени.

R-I-операции выполняются с помощью сенсоров и могут быть определены как "сенсорные" операции, а *I-R*- операции выполняются с помощью ефекторов (исполняющих устройств) и могут быть определены как "ефекторные" операции. Таким образом, в природе с помощью коммуникатов и объектов-оперантов могут иметь место технологические операции и процессы отображения некоторой сущности материальных объектов в информационные и, наоборот, информационных объектов в материальные.

Особенности R-I-преобразований. *R-I*-операции более широкое понятие, чем общепринятые «операции ввода – вывода». В информатике, *ввод/вывод* трактуется как взаимодействие между обработчиком информации (например, компьютер) и внешним миром, который может представлять как человек, так и любая другая система обработки информации. В нашем случае R-I-операции включают также взаимодействие коммуниката и коммуниканта. Технически технология преобразования материального объекта в его образ может быть реализована двумя способами:

- *Анализ с помощью технического средства* (соответствующего сенсора) *первичного коммуниката δ (*например, электромагнитного излучения или вещественного потока), которые создаются и эммитируются коммуникатором (рассматриваемого как активный объект);
- *Зондирование* коммуникатора некоторым влияющим коммуникатом *λ* (который бомбардирует пассивный объект - коммуникатор) и анализ с помощью соответствующего сенсора вторичного коммуниката *δ*, (который еммитируется коммуникатором*).*

I-I- ***преобразование*** – известная основная операция теории связи, например, передача изображения из одного компьютера на другой; трансакция в ИС.

Концепция отображения универсума

Объектологический подход к представлению макромира. С точки зрения "наивного реалиста", позиции которого будем разделять в данной статье, Вселенную и ПдО можно рассматривать как некоторые сетевые структуры, в узлах которых находятся взаимодействующие между собою объекты.

Непосредственные взаимодействия. Современные экспериментальные данные свидетельствуют, что существует, строго говоря, только четыре качественно разных вида взаимодействий: гравитационное, электромагнитное, сильное, слабое. Эти взаимодействия считаются фундаментальными, т.е. самыми основными, исходными, первичными. Эффект взаимодействий существенно зависит от размеров объектов и расстояния между ними. При взаимодействии материального объекта r_1 с другим материальным объектом r_2 результатом всех видов взаимодействий может быть как формирование (синтез) новых объектов $(r_3\text{-}r_n)$ (вещественный процесс), так и «отражение» - изменение состояния одного из взаимодействующих объектов таким образом, что измененное состояние одного (например, r_2), отображает некоторые свойства другого объекта (r_1). Последний можно рассматривать как информационный процесс – процесс формирования некоторого информационного образа i материального объекта r_{in}

Опосредствованные взаимодействия. В [8] показано, что в соответствии с ролью и природой действий одних материальных объектов на другие взаимодействие *актантов* может быть отображено в каморках, соединенных в соответствии со схемой на рис.3. Многозначность понятия сигнал нарушает строгость описания информационных процессов и не позволяет его использовать в случае, например, описания обычных процессов почтовой связи. Будем пользоваться такими метаонтологическими определениями актантов: **Коммуникант** - участник коммуникации, задействованный в коммуникативном акте взаимодействия между коммуникатором и реципиентом. ***Коммуникатор*** - участник взаимодействия, которое может порождать коммуникаты – вещественные, энергетические или полевые потоки, в общем случае потоки любой природы (материальные, денежные, информационные). ***Коммуникат*** - вещественный, энергетический или полевой потоки, которые способны осуществлять материальное или информационное влияние на реципиента. ***Реципиенты*** – характеризуются состоянием и совокупностью реакций (функцией). ***Операнты*** - материальные объекты, которые обеспечивают опосредствованное воздействие объектов - коммуникаторов на объекты-реципиенты; реализуются в виде машин, людей, человеко-машинных систем или систем, которые могут содержать и естественные объекты. ***Объекты-инфлюанты*** - управляющие, координирующие или иным образом влияющие на объекты-операнты и ими обусловленные процессы. **Среда** - то, что окружает систему и оказывает на нее влияние; может быть внешней и внутренней.

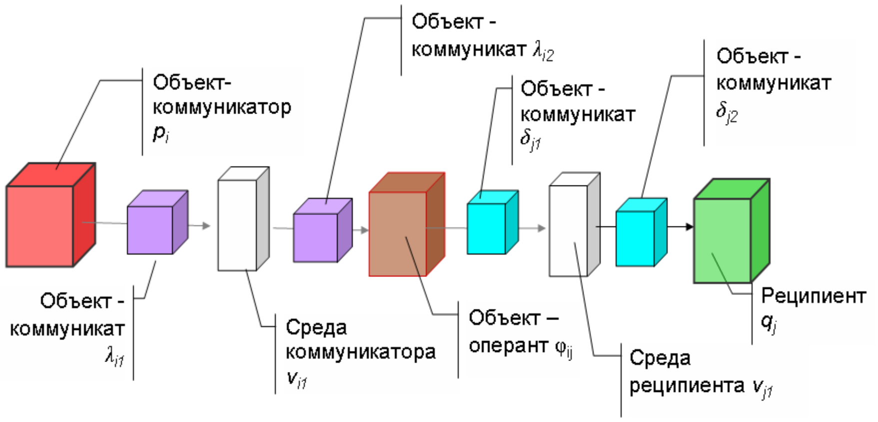


Рис.3 - Схема соединений каморок для отображения действия коммуникатора на реципиента

Как отмечалось, объекты ПдО могут образовывать сложные сетевые структуры. По аналогии с отображением свойств объектов в базе данных, предложенным Кодом, предлагаемая концепция построения модели исходит из возможности отображения сущностей объектов ПдО в виде композиции *каморок*, по существу представляющих некоторые носители отображений соответствующих объектов универсума. Для отображения объектов и взаимодействий каморки могут быть двух форматов (рис. 4) . Каморки могут содержать поверхностные ячейки (из двумерных пикселей), объемные (воксели) и объемно-пространственные ячейки (имеющие внутреннюю структуру) и допускать запись в них и считывание ссылок к внешней памяти.

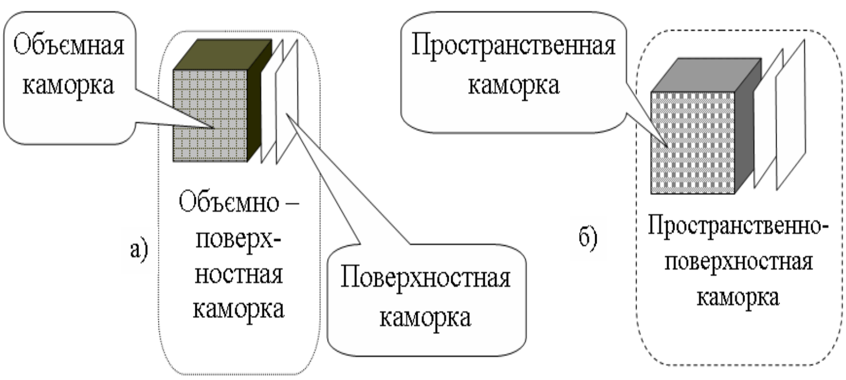


Рис. 4 Форматы каморок

При наличии *n* классов объектов *R* количество вариантов связей классов возрастает пропорционально n^2, поэтому модель на рис. 3 становится ненаглядной. Кроме того, иногда важным есть взаимодействие с объектами окружения ПдО - миром внешней действительности.

Отображение множества материальных и информационных объектов

Предложенная в [8¸9] модель универсума на основе композиции каморок отображения позволяет отобразить несколько категорий и классов множества объектов и совместить таким образом физический и онтологический принципы отображения ПдО. Модель может быть реализована в материальных или виртуальных средах.

Принцип построения каморковой модели универсума состоит в выделении для каждого объекта u_i универсума *U* каморки объекта d_i из множества *D*, формировании в каморке объекта d_i отображения сущности объекта или ссылки на внешнюю память, объединению каморок, упорядоченных по категориям и ролевым признакам отображаемых объектов, в структурные элементы - модули, например, в форме параллелепипедов, и формировании отображения универсума объектов в виде композиции модулей D_α , где *α*- - множество индексов модулей.

На Киевской сессии KDS-2010 в докладе [8], а также в [9] была представлена «модель универсума» как композиция каморок в виде несимметричного «кубуса» (фото 1) у левой грани которого расположены каморки коммуникаторов, а у верхней грани – каморки реципиентов (12 категорий). На фиг. 5 приведен пример графики проекции композиции каморок на грань коммуникаторов. На грани представлены следующие категории: Время, Среда, Естественные объекты, Искусственные объекты, Биологические, Системы, Образы сущностей материального мира, элементы и конструкции информационного мира, Продукты деятельности на неживых носителях, ментальные объекты, не обнаруживаемые объекты и языки. В настоящей статье, являющейся идеологически продолжением предыдущей, предлагается, на наш взгляд, более последовательное изложение концепции построения кубуса и отличающаяся композиция каморок. Композиция отличается введением некоторой симметрии, а также большим количеством категорий - 16, которые нанесены на гранях кубуса (при заданных высоте строк и размерах шрифта).

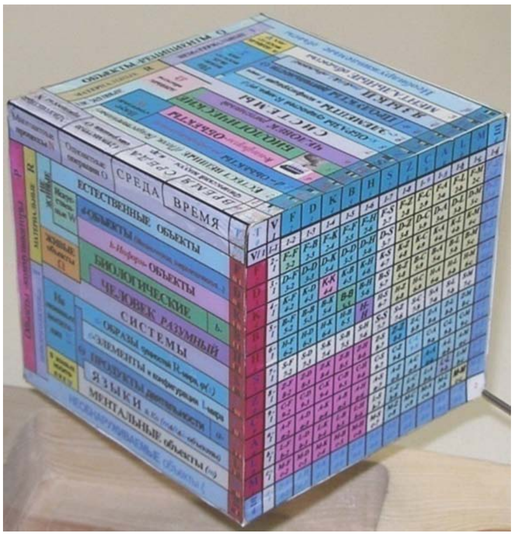


Фото 1

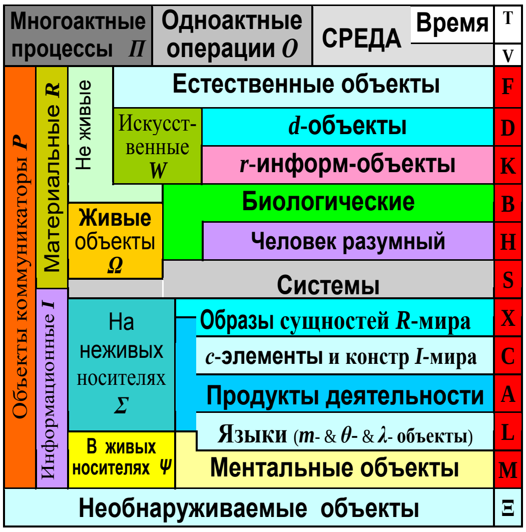
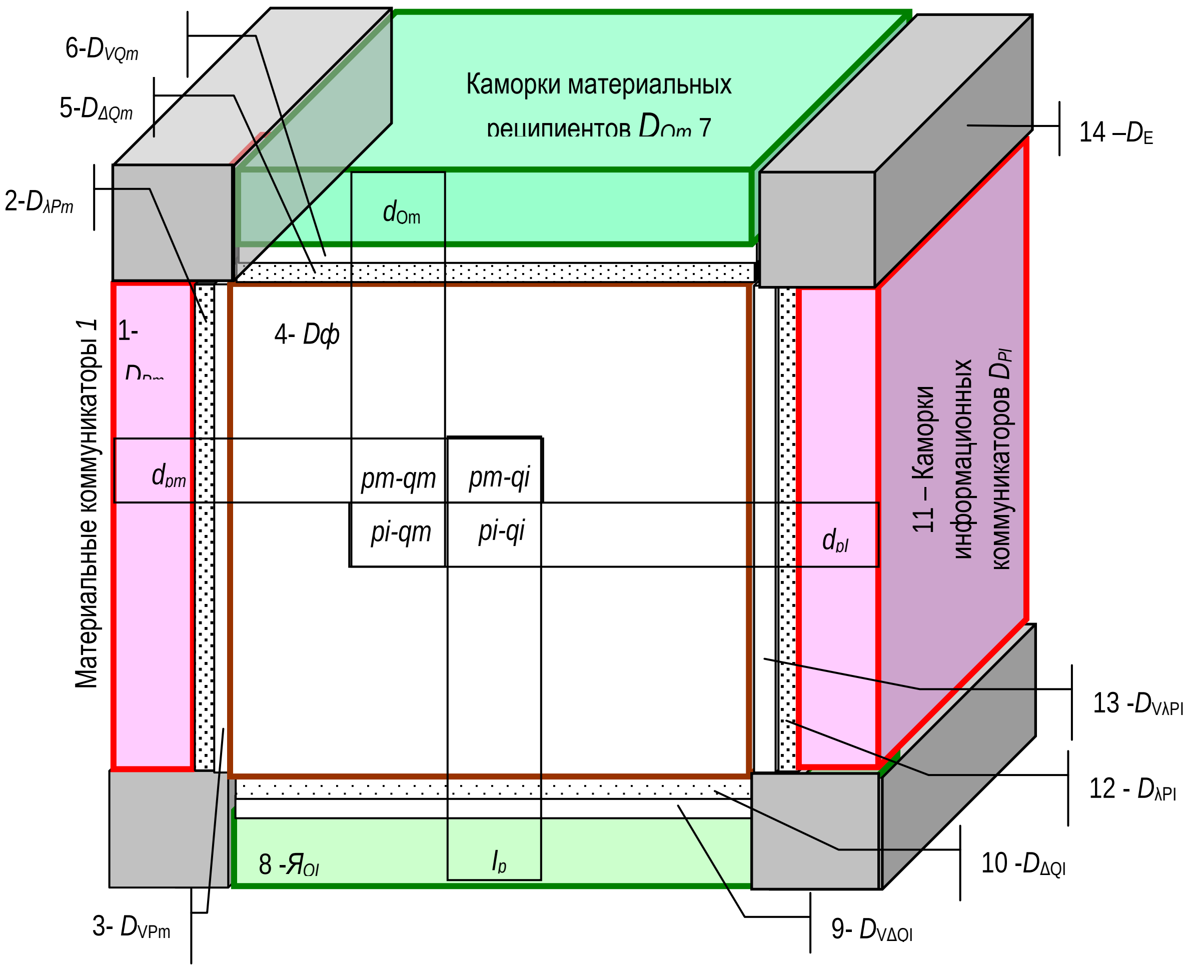


Рис. 5 Базовые категории универсума

Формализация отображения ПдО. В основу онтологического описания взаимодействующих на физическом уровне объектов ПдО положены структурные элементы схемы на рис. 3. При таком подходе *множество материальных объектов R* ПО формально может быть представлено пятеркой *R*= < *P, Q, Λ, Δ, Φ*>, где *P* - множество объектов-коммуникаторов, которые являются источниками формирования множества объектов-коммуникатов *Λ; Q* - множество объектов-реципиентов, на которые воздействуют объекты-коммуникаты *Δ; Ф* - множество объектов-оперантов, которые принимают участие в формировании вещественных или полевых действий на объекты-реципиенты из множества *Q. Объекты-коммуникаты Λ, Δ* - есть носители воздействий: это вещественные потоки *G* (характеризуются массой покоя) и физические поля *Г,* которые в общем случае могут характеризоваться потенциальной и кинетической энергией *Е.* По аналогии с известной моделью «сущность отношение» (ER-модель) конструктивный компонент предлагаемой модели может быть представлен в виде: триады сущностей ***E*** (коммуникатор, оперант и реципиент) и взаимодействий *i* (impact *v* interaction) между сущностями, реализуемые через коммуникаты *i* (исходящий и воздействующий коммуникаты). Этот компонент может быть принят за базовый и, по аналогии с ER-моделью, может быть обозначен, как ***E-i-E-i-E***- модель. Для учета сущности, управляющей воздействием на оперант, компонента может быть соответственно усложнена и обозначена как ***E-i-***^{E}***E-i-E*** – модель.

Композиция модели отображения в форме куба

Расположение модулей. Каморковая модель отображения универсума выполнена согласно рис. 6 в форме куба в виде композиции семнадцати модулей. Расположение модулей привязано к правой системе Декартовых координат. (



Фиг. 6 Структура метамодели универсума в форме куба с раздельными модулями материальных и информационных коммуникантов

Измерения объектов визуализируются в направлениях осей Х,У и Z, а именно: У - ось измерений категорий коммуникаторов; Х - ось измерений категорий реципиентов; Z - ось измерений классов коммуникантов. m, n - соответственно объекты или классы взаимодействующих коммуникаторов и реципиентов. Члены измерений визуализируются как точки или звенья, которые откладывают на осях модуля.

Обозначения каморок: *D-Den; P-* объекты-коммуникаторы, *Q-* объекты-реципиенты, P_m - материальные объекты коммуникаторы, Q_r – материальные объекты реципиенты, P_i – информационные объекты коммуникаторы, Q_i – информационные объекты реципиенты, λ_{km} – коммуникаты материальных коммуникаторов, λ_{ki} – коммуникаты информационных коммуникаторов, δ_{rm} – коммуникаты материальных реципиентов, δ_{ri} – коммуникаты информационных реципиентов, *Ф* – объекты-операнты.

Состав и функциональное назначение модулей следуюшие:

Модуль 1 (III октанта) содержит каморки материальных объектов-коммуникаторов множества (D_{pm}). Графические элементы модуля 1 на грани 1 (рис. 6, фото 2) расположены, таким образом, что последовательность категорий объектов-коммуникаторов направлена по оси -У, а в направлении оси Z отображаются классы, подклассы, абстрактные и/или конкретные объекты-коммуникаторы.

Модуль 2 (III октанта) содержит каморки $D_{\Lambda Pm}$ исходящих коммуникатов *P*. Примеры коммуникатов: фотонный поток, электронный поток, радиоволна (в общем случае электромагнитная волна), лист бумаги, конверт, посылка почтовая или контейнер транспортных перевозок.

Модуль 3 (III октанта) содержит каморки $D_{V\Lambda}$ для отображения пространства объектов среды, в котором распространяются исходные коммуникаты *Λ* между коммуникаторами *Pm* и входными элементами объектов-оперантов *Ф*.

Модуль 4 (IV октанта) содержит каморки $D_Ф$ для отображения объектов-оперантов $U_ф$, с помощью которых осуществляется взаимодействие объектов-коммуникаторов *P* и объектов-реципиентов *Q*. Сущность объектов-оперантов состоит в превращении материальной $(g_\lambda, \gamma_\lambda)$ и энергетической e_λ компонент коммуниката *Λ* в материальные $(g_\delta, \gamma_\delta)$ и/или энергетические (e_λ) компоненты коммуниката *Δ*. Учет возможности на входах и выходах разных комбинаций *g-,γ--* и *e*-компонент коммуникатов в общем случае создает достаточно много вариантов (64) превращений коммуникатов. В практике конкретные вещественные *G,* полевые *Г* или энергетические *E* составляющие или их комбинации есть доминирующие, а другие могут быть неактуальными.

С учетом этого для некоторых вибраных входа и выхода операнта, в простейшем случае, можно выделить такие категории *<u>операций</u>* объектов-оперантов (в отличие от операций объектов коммуникантов:

r-r - операции, сущность которых определяется превращением материальной $(g_\lambda, \gamma_\lambda)$ и/или энергетической e_λ составных коммуниката *Λ* в материальные $(g_\delta, \gamma_\delta)$ и/или энергетические (e_λ) компоненты коммуниката *Δ*. Операции синтеза и декомпозиции сложных операций *r-r* категории могут быть описаны методами формальных технологий [10].

i-i - операции, в которых входной информационный объект множества *I* (или поток объектов), преобразовывается в выходной также информационный объект (или поток информационных объектов). Операции выполняются материальными объектами-оперантами (аппаратными, программно-аппаратными, программными, а также биообъектами, которые могут создавать информационные среды); на верхнем уровне обычно описываются логико-математическими моделями, которые не включают физические преобразования материальных носителей информационных объектов.

i-r операции, в которых исходный информационный объект множества *I* (или поток объектов), независимо от носителя, в конечном счете опрдляет выходной материальный и/или энергетический объект (или поток объектов). Операции выполняются исполнительными устройствами (ефекторами, которые с помощью драйвера исходного устройства реализуют действие информационного объекта на материальный объект или систему и агрегатами в более сложных случаях).

Трактовка *r-i* операций - очевидна. В простейшем случае выполняются сенсорами (формирование образов сущностей материальных объектов).

Информационные среды, в которых выполнение *i-i, r-i й i-r* операций реализуется в виде искусственных твердых конструкций, в практике называются «аппаратная часть». Для общности, например, со случаем использования программно-аппаратных средств или радиоволн в ионосфере, *i-i, r-i и i-r* преобразователи, при рассмотрении их в роли объектов-коммуникантов, в данной работе имеют обобщенное название *r*-информ-объекты или *ri*-объекты.

Модуль 5 (I октант) содержит каморки действующих коммуникатов $D_{\Delta Qm}$. В этих каморках отображаются объекты-коммуникаты Δ, которые воздействуют на объекты-реципиенты Q_m.

Модуль 6 (I октант) содержит каморки D_{VQm} для отображения пространства объектов среды между исходными элементами объектов-оперантов Φ и реципиентов Q_m, в котором распространяются действующие коммуникаты Δ_{Qm}, (в простейшем случае содержит каморки пространственных интервалов).

Модуль 7 (I октант) содержит каморки объектов-реципиентов D_{Qm}. Модуль 7 расположен ортогонально модулю 1 и графические элементы грани 5 расположены так, что подмодули каморок последовательности категорий объектов-реципиентов направлены по оси X, а классы, подклассы и конкретные объекты-реципиенты могут отображаться в каморках в направлении оси Z.

Модуль 8 (IV октант) содержит каморки объектов-реципиентов D_{Qi}. Модуль 8 расположен ортогонально модулю 1 и графические элементы грани 5 расположены так, что подмодули каморок последовательности категорий объектов-реципиентов направлены по оси X, а классы, подклассы и конкретные объекты-реципиенты могут отображаться в каморках в направлении оси Z.

Модуль 9 (IV октант) содержит каморки D_{VQi} для отображения пространства объектов среды между исходными элементами объектов-оперантов Φ и реципиентов Q_i, в котором распространяются действующие коммуникаты Δ_{Qi}.

Модуль 10 (IV октант) содержит каморки действующих коммуникатов $D_{\Delta Qi}$. В этих каморках отображаются объекты-коммуникаты Δ, которые воздействуют на объекты-реципиенты Qi.

Модуль 11 (I октант) содержит каморки информационных объектов-коммуникаторов множества (D_{pi}). Графические элементы модуля 1 на грани 1 (рис. 7, фото 2) расположены, таким образом, что последовательность категорий объектов-коммуникаторов направлена по оси -У, а в направлении оси Z отображаются классы, подклассы, абстрактные и/или конкретные объекты-коммуникатори.

Модуль 12 (I октанта) содержит каморки $D_{\Lambda Pi}$ исходящих коммуникатов P_i. Примеры коммуникатов аналогичны коммуникатам материальных коммуникаторов: фотонный поток, электронный поток, радиоволна (в общем случае электромагнитная волна), лист бумаги, конверт, посылка или контейнер транспортных перевозок.

Модуль 13 (I октант) содержит каморки $D_{V\Delta}$ для отображения пространства объектов среды, в котором распространяются исходные коммуникаты Δ между коммуникаторами Pi и входными элементами объектов-оперантов Φ.

Модуль 14 – содержит каморки описание среды и мер времени. 15 – содержит изложение метатеории и описание конкретного кубуса. 16 – содержит описание конкретных процессов и операций, созданное пользователем. 17 – содержит описание источника и описание команд, множество потоков команд, координационных воздействий.

Описание графики граней физической реализации модели

На фото 2 приведен пример физической реализации модели универсума в виде куба. Графические элементы поверхностных ячеек на гранях модели определяется проекцией на грани соответствующих подмножеств каморок объектов-актантов. Эти проекции приведены на развертке поверхностей граней

1,2,3 (рис. 7) в виде областей, на которых нанесены категории универсума и, при необходимости, иллюстрации к ним.

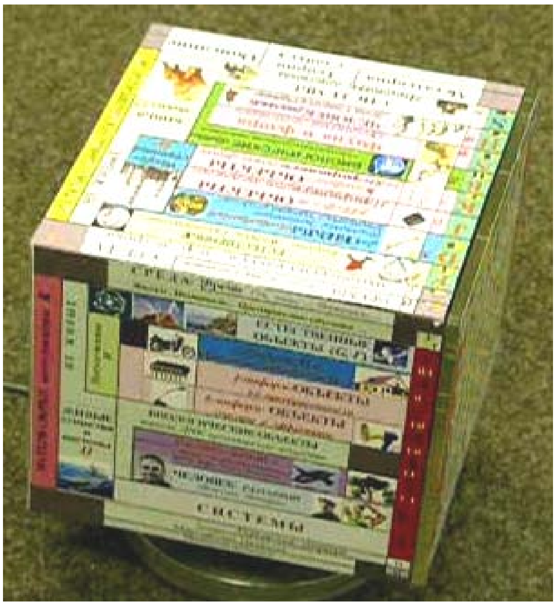


Фото 2 Общий вид модели с разделенными модулями материальных и информационных коммуникантов в форме куба

Множество объектов *R* наряду с подмножествами неживых объектов *F*, живых - *Ω* и искусственных объектов *W* также включают подмножество систем *S*, которые могут представлять определенную совокупность конфигураций элементов разных категорий. Живые объекты *Ω* - могут рассматриваться на разных уровнях: молекулярном, субклеточном, клеточном, органно-системном, организменном и других. Поэтому этот термин достаточно широкий и не однозначный. Объем понятия включает подмножество множества биологических объектов *B*. Биологические объекты *B* включают категорию «человек» *H*. Искусственные объекты *W* представлены категорией не информатизированых объектов *D* (*d*-объектов) и категорией *K* (*r* -информ-объектов), которые могут оперировать также с информационными объектами категории *I*.

В множестве информационных объектов *I* выделено подмножество ментальных объектов *Ψ* и подмножества информационных объектов на неживых носителях *Σ. Ψ* - подмножество ментальных объектов, порожденных в индивидууме или коллективе, живых организмах в результате познавательной деятельности (образы внешнего мира, идеальные объекты (эмпирические и теоретические (конструкты, абстрактные объекты)), сигналы, конфигурации, мысли, базовые когнитивные операции, ...). Подмножество *Σ* (информационные объекты на неживых носителях) включает подмножество *X* - образы сущностей *R*-мира и подмножество *A* - продукты деятельности информационного мира . Категория объектов X включает множество X_1 - информационных объектов, которые отображают (в виде абстрактных или конкретных моделей либо иллюстраций) некоторые сущности *R*-мира, в котором *x'*-образы, сформированные материальными средствами без участия оператора, и множество *X2,* в котором *x''*-образы, представляющие *x'*-образы, однако скорректированы действующим индивидуумом, например, фотопортрет, отретушированный художником соответственно его мировосприятию *(*у= φ( )) или формализованную модель некоторой сущности, виртуальный объект.

В категорию *Θ* - продукты деятельности информационного мира объединены любые медиа-продукти умственной деятельности и зафиксированные на носителях (языка, тексты, компьютерные программы, рисунки, мелодии, фильмы, устное творчество, ...), которые доступны пользователям. Множество *Продукты деятельности I-* мира включает *c*-элементы и собранные из них генетические конструкции *I*-мира, которые могут использоваться для построения более сложных конфигураций. Элементы категории *C* (*c*-элементы) представляют первичные математические понятия (точка, линия, поверхность, объемная фигура, *n*-мерная фигура, число, множество, математические структуры) или отображение простейших физических сущностей материального мира (в виде точечных, линейных, объемных элементов для макромира и пространственных структур микромира - атомов, частиц и т.д.), которые моделируют свойства вещественных объектов и физических полей.

В множестве информационных объектов *I* выделена специфическая категория отображения объектов *U:* язык *L.* Кроме ментальной составляющей *m,* составляющей на неживых носителях (алфавит,...) *θ* включает также коммуникационную составляющую *λ* в виде образцов звуковых колебаний в некоторой материальной среде и их отображении на носителе, например магнитной пленке.

Категория объектов, которые не обнаруживаются инструментальными средствами, обозначены как *Ξ (*ангелы, Бог, русалки, домовой, НЛО).

На грани 2 (рис. 7) нанесена категориальная сетка *каморок объектов-оперантов* в форме квадратной матрицы, поля каморок исходных и действующих коммуникатов (узкие горизонтальная и вертикальная полосы), мнемонические обозначения категорий *объектов-коммуникантов.* Мнемоника обозначения классов *объектов-оперантов* образованная путем соединения *i*-го мнемонического обозначения основного *объекта-коммуникатора* с *j*-м обозначением основного *объекта-реципиента* (рис. 7). При многообъектных взаимодействиях мнемоника соответственно усложняется.

Рис. 7 Графика граней 1,2 и 3 кубуса

1) грань материальных коммуникаторов

2) грань информационных коммуникаторов

3) грань оперантов

Грани 3,4 и 6 могут использоваться для иллюстраций характерных образов категорий объектов.

Отметим, что для упрощения возможности взаимодействия оператора с компьютером оборудованным двумерным монитором, вместо трехмерного отображения универсума может быть использовано полученное из него двумерное (рис. 8) в виде развертки граней 1,2 3, 5 и 6 куба.

Формирование траектории развития

Среда, в котором происходят протяженные процессы взаимодействия объектов, движение объектов в пространстве *V* и времени *T*, переходные характеристики процессов изменения состояния объектов, их метрика отображаются в каморках $Я_{СТП}$ модуля пространственно-временных интервалов и процессов 14.

Возможные изменения структуры и динамика ПО при процессах, которые содержат более, чем один акт взаимодействий, могут быть отображены, например, в виде *n*-мерной матрицы. На рис. 8 – двумерный случай, в котором: C_{11} – отображение актуального состояния объектов либо некоторой сетевой структуры; C_{12}, C_{21}… Cnn – возможные траектории развития процесса. Развертка, как на рис. 8, может иллюстрировать формирование структуры «обликов» объектов и соответствующего пространства, в общем случае динамики физических процессов как образа бытия. На основе развертки могут строиться каузально-эмпирические модели, n-мерные модели взаимодействий верхнего уровня для уже конкретных процессов (концептуальные, графы-модели, логические) и модели управления, определяться такие понятия, как терминальные объекты, такты, агрегации измерений, действительность, методы ИИ. При огромной последовательности операций или процессов модели могут отображать переход к предельному случаю - синэргетическим процессам. Таким образом, на основе развертки воздействий коммуникантов могут быть сделаны эмпирические обобщения для конкретной ПдО и продемонстрирован переход от физических знаний к логическим. Схема может использоваться для демонстрации процессов развития на уровне идей (дао-траектории в китайском даосизме, реинкарнаций индийской философии и др.).

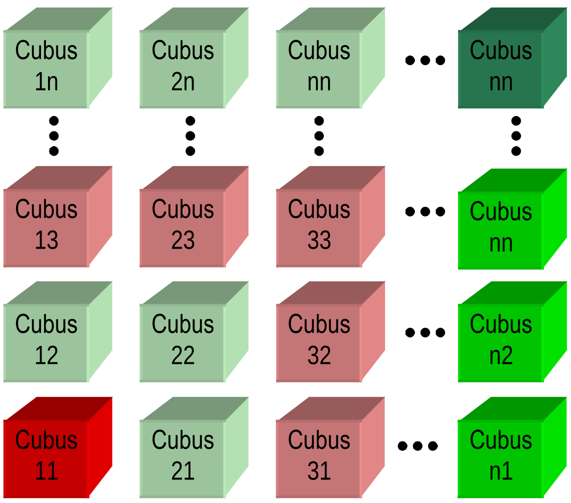


Рис. 8 Двунаправленное развитие состояний кубуса

Выводы

На основе важнейших поведенческих свойств материальных объектов - способности обмениваться с внешней средой веществом *G*, излучением, энергией *E* и импульсом – введены понятия «коммуникат», «оперант», «инфлюант», которые позволили визуализировать взаимосвязи объектов универсума на физическом уровне и важнейший процесс бытия: отражение свойств объектов универсума в состояниях других объектов описать с помощью коммуникатов .

Таким образом физико-онтологическая модель позволяет решить¸либо приблизить решение следующих задач:

- Объединение различных картин мира и разработка концепции целостной картины мира;
- Гносеологическая задача:: создание инструментального средства познания;
- Разработка базы знаний для обеспечения практического функционирования ИИ ;
- Концентрированное отображение сетевой структуры мира;
- Разработка концепции интерфейса базы знаний;
- Инициация мышления при не сформулированных проблемах.

Разработана каморковая модель отображающая категории и природу материальных и информационных взаимодействий объектов. Модель может быть выполнена в виде трехмерной физической модели, позволяющей хранить наиболее обобщенные наработанные человечеством знания о мире (как наглядное пособие для обучения принципам классификации, устройство ввода, сувенир), а также более глубокие знания - в виде программной модели компьютера. Возможное применение компьютерной модели - в составе агрегата «обобщенная компьютерная модель универсума - конкретная модель ПдО» для хранения, идентификации и классификации объектов в информационных системах, работающих со знаниями.

Благодарности.

"The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine (www.aduis.com.ua)."

Ссылки

[Кургаев , 2008] Кургаев А.Ф. Проблемная ориентация архитектуры компьютерных систем.- Киев: Сталь, 2008.- 540 с.

[Боровиков, 2002] С. Е. Боровиков, С. Б. Переслегин О понятии развития в симметричной мета-онтологической картине мира http://www.igstab.ru/materials/Pereslegin/Per_Symetrik.htm

[Степин, 2004] Степин В.С. Научная картина мира в культуре техногенной цивилизации / В.С.Степин, Л.Ф.Кузнецова. - М.: Изд-во Проспект, 2004.- С.18.

[Палагин, 2009] . Палагин А.В., Петренко Н.Г.Системно-онтологический анализ предметной области// УСиМ – 2009. № 4. – С.3– 14.

[Лозовский, 2003] Лозовский В.С. К десигнативной теории имен. http://vloz.mylivepage.com/wiki/155/83

[Палагин, 2009] Александ Палагин , Андрей Митхайлюк, Виталий Величко, Николай Петренко. К интеграции онтологий предметных областей. XVI - International conference “Knowledge – Dialogue – Solution” KDS-2010, Kuiv, Ukraine, September, 2010.

[Переслегин , 2001] С.Б. Переслегин, Информационный объект http://traditio.ru/wiki/ .

[Мержвинский, 2010] Мержвинский А.А. Модель унмиверсума Informations Models of Knowledge, XVI-th International Conference Knowledge – Dialogue – Solution, N. 15. – ITHEA, Sofia, 2009. p.31- 39.

[Мержвинский , 2009]. Патент Промышленный образец № 19543 від 12.10.2009 «КОМПЛЕКТ ПРИСТРОЇВ ДЛЯ ВІДОБРАЖЕННЯ УНІВЕРСУМА». Автор: Мержвинський Анатолій Олександрович.
http://base.ukrpatent.org/searchBul/search.php?action=viewdetails&IdClaim=21733&chapter=variants&dbname=pp

[Крылов, 1997]. Крылов С.М. Формальная технология в философии, технике, биоэволюции и социологии.- Самара: СамГТУ, 1997.-180 с.

Authors' Information

Мержвинский Анатолий Александрович *– Институт кибернетики им. В.М.Глушкова НАНУ, 03680 МСП Київ-187, пр-т Академіка Глушкова,40, Украина;*
e-mail: merjv@mail.ru
Major Fields of Scientific Research: физико-технологические проблемы кибернетики, микро-оптоэлектроника, биосенсорика

ИНТЕГРАЦИЯ ГЕТЕРОГЕННЫХ ИСТОЧНИКОВ ДАННЫХ НА ОСНОВЕ РЕКУРСИВНОЙ ДЕКОМПОЗИЦИИ

Алексей Кашников, Людмила Лядова

Abstract: Управление данными на современном предприятии характеризуется наличием большого количества разнородных источников данных, не связанных едиными механизмами управления, в том числе и слабоструктурированных или неструктурированных данных и т.п. При этом модель данных, лежащая в основе большинства систем, – реляционная – не является эффективной для решения многих задач. Раздельно существуют системы аналитической обработки и оперативного управления данными, системы управления документами и пр. Различные задачи требуют использования различных моделей представления данных. На этом фоне ставится задача интеграции гетерогенных данных, эффективное решение которой требует создания модели интеграции (или «интеграции» различных моделей данных), которую можно было бы рассматривать как основу для реализации системы, поддерживающей оперативное управление разнородными данными и их аналитическую обработку. Предлагается модель интеграции данных, в которой должны поддерживаться унифицированное представление разнородных источников данных, управление ограничениями целостности, управление выполнением операций манипулирования данными и запросов, согласование данных из разных источников, возможность расширения и настройки на новые источники данных. Существующие подходы к интеграции гетерогенных данных имеют ограничения, которые не позволяют в полной мере говорить об их универсальности. Предлагаемый подход к интеграции основан на рекурсивной декомпозиции источников данных, при которой каждый источник данных последовательно разбивается на атомарные элементы данных, причем на каждом уровне рекурсивной вложенности данные и их описания представляются единообразно. Такая модель позволяет осуществлять интеграцию различных источников данных на любом уровне посредством задания связей между произвольными элементами схемы, ограничений целостности и допустимых операций. Разработанная модель представления источников данных, которая позволяет осуществлять многоуровневую интеграцию гетерогенных источников данных в единое информационное пространство, обеспечивает поддержку ограничений целостности на любом уровне интеграции источников данных, а также поддержку структурных и ассоциативных связей между источниками данных на любом уровне интеграции. Обеспечивается возможность динамического изменения схемы данных, а также расширяемость системы за счет новых моделей данных.

Keywords: модель данных, гетерогенные системы, интеграция.

ACM Classification Keywords: H. Information Systems: H.2. DATABASE MANAGEMENT (E.5): H.2.1. Logical Design – Data models.

Введение

Активное развитие информационной инфраструктуры предприятий привело к возникновению проблем, препятствующих эффективному использованию имеющихся данных и, как следствие, принятию качественных управленческих решений:

Большое количество разнородных источников данных с различными механизмами управления.

Реляционная модель в качестве «универсальной» основы управления данными. Между тем, реляционная модель данных не является адекватной для решения многих современных задач – различные задачи требуют различных моделей представления данных.

«Излишняя» функциональность современных систем управления базами данных (БД) и, как следствие, излишняя дороговизна. Современные коммерческие СУБД предоставляют огромный объем функциональности, большая часть которой остается невостребованной.

Большой объем данных, значительная часть которых представлена слабоструктурированными или неструктурированными данными.

Раздельное существование систем аналитической обработки данных и систем оперативного управления данными.

Единое информационное пространство является ключевым фактором успешности современного бизнеса. В Клермонтском отчете [1] говорится, что одной из важнейших целей сообщества баз данных является переход от управления традиционными БД к задаче управления наборами структурированных, полуструктурированных и неструктурированных данных, распределенных по многим репозиториям предприятий и узлам Internet. Как следствие, возникает задача интеграции гетерогенных данных, эффективное решение которой требует создания *модели интеграции* (или «интеграции» различных моделей данных).

Общей целью данной работы является разработка *системы интеграции гетерогенных источников данных*. В данной статье рассматривается задача разработки *модели интеграции данных*, в которой должны быть отражены следующие аспекты:

- унифицированное представление источников данных;
- управление ограничениями целостности;
- управление выполнением операций манипулирования данными и запросами;
- согласование данных из разнородных источников;
- возможность расширения и настройки на новые источники данных.

Существующие подходы к интеграции гетерогенных данных имеют следующие особенности, которые не позволяют в полной мере говорить об их универсальности:

Вопросы интеграции гетерогенных данных рассматриваются преимущественно с позиций осуществления унифицированных запросов, а не комплексного управления данными.

Манипулирование данными рассматривается преимущественно при интеграции только реляционных СУБД (либо в случае, если нереляционная СУБД имеет реляционную оболочку).

В качестве единой модели данных рассматривается реляционная схема, либо в более общем случае – табличное представление данных.

Для достижения поставленной цели должны быть решены также следующие задачи:

разработка языка запросов и манипулирования интегрированными данными;

разработка программной модели и архитектуры системы интеграции гетерогенных источников данных;

разработка механизмов отображения существующих источников данных на интегрированную модель.

Рассматриваемый в данной работе подход к интеграции основан на *рекурсивной декомпозиции источников данных*, при которой каждый источник данных сводится к неделимым элементам данных, причем на каждом уровне рекурсивной вложенности данные и их описания представляются единообразно. Такая модель позволяет осуществлять интеграцию различных источников данных на любом уровне декомпозиции посредством задания связей между произвольными элементами схемы, ограничений целостности и допустимых операций.

Гетерогенные информационные системы: подходы к интеграции данных

Необходимость в интеграции гетерогенных данных возникает в различных условиях и требует различных подходов, соответствующих этим условиям и требованиям, которые предъявляются к информационным системам. В некоторых случаях достаточно обеспечить работающие информационные системы шлюзами для обмена данными, в других – обеспечить единое представление информационного пространства для возможности выполнения запросов, охватывающих различные источники данных, в третьих – необходимо предоставить комплексный инструментарий по управлению данными в гетерогенной среде, включая выполнение транзакций и поддержку ограничений целостности. Более глубокая интеграция требует более сложного решения специфических проблем, обусловленных гетерогенной средой. Анализ различных аспектов функционирования гетерогенных систем демонстрирует необходимость применения иных подходов к управлению данными, чем в случае разработки традиционных систем. Разработчики информационных систем (ИС), часто сталкиваясь с набором слабо связанных источников данных, вынуждены каждый раз решать низкоуровневые задачи управления данными в разнородных коллекциях. В число этих задач входят обеспечение возможностей поиска и запрашивания данных; соблюдение правил, ограничений целостности, соглашений об именовании и т.д.; отслеживание происхождения данных; обеспечение доступности, восстановления и контроля доступа; управляемое развитие данных и метаданных и пр. Гетерогенность данных разделяется на физическую и семантическую [16].

Физическая гетерогенность подразумевает различия в представлении данных; может выражаться в различии типов данных, реализаций моделей данных (например, различные реляционные системы). Могут различаться также языки описания процедур, триггеров, языки запросов, манипулирования и определения данных. Кроме этого, могут различаться и сами модели данных. Данные могут быть представлены не только в БД, но и в электронных таблицах или почтовых файлах и т.д. Проблема физической гетерогенности решается посредством *введения стандартов взаимодействия* (ODBC, DAO, OLE DB, ADO, ADO.NET). Достоинством универсальных механизмов является возможность применения одних и тех же средств доступа к разным типам источников, поэтому приложения легко модифицировать, если необходима замена СУБД. Но за универсальность приходится платить невозможностью доступа к уникальной функциональности, специфичной для конкретной СУБД, снижением производительности.

Семантическая гетерогенность проявляется в различиях в наименованиях данных, значениях и логических структурах [17]. В данной области проводится большое количество исследований, но до сих пор не появилось какого-либо стандарта. Прототипы реляционных языков (MSQL, IDEAL) были разработаны для решения этой проблемы. Многие специальные функции этих языков нуждаются в выразительной мощности, выходящей за пределы существующих реляционных систем, диалектов SQL. Требуются мультиреляционные операции и логика исчисления предикатов более высокого порядка, тогда как SQL основан на реляционной алгебре и логике предикатов первого порядка.

Гетерогенная система должна иметь собственные *механизмы управления целостностью*. Каждая локальная БД может иметь свои механизмы поддержания целостности, но контроль целостности на глобальном уровне необходим, так как на нем могут появиться новые связи и новые ограничения. Еще одна задача – поддержка *транзакций*. В гетерогенной среде концепция глобальной транзакции, состоящей из строго определенных субтранзакций, может быть слишком строгой.

В качестве новой абстракции управления данными в таких сценариях вводится понятие пространства данных [12]. *Пространство данных* должно содержать всю информацию, необходимую конкретной организации, несмотря на формат и местоположение этой информации, а также моделировать развитый набор связей между репозиториями данных. Пространство данных моделируется как набор участников и связей. *Участниками пространства данных* являются индивидуальные источники данных: они могут быть реляционными БД, репозиториями XML, текстовыми БД, Web-сервисами и пакетами программного обеспечения. Они могут храниться или быть потоками данных (локально управляемыми системами потоков данных) или даже сенсорными установками. Некоторые участники могут поддерживать выразительные языки запросов, а другие – быть неинтеллектуальными и поддерживающими лишь ограниченные интерфейсы для формулировки запросов (структурированные файлы, Web-сервисы и пр.). Участники могут быть структурированными (например, реляционными БД), полуструктурированными (XML, коллекции кода) или полностью неструктурированными. Некоторые источники будут поддерживать операции обновления, другие – допускать только добавление (в целях архивации), третьи – не допускать изменений вообще. Пространство данных должно *моделировать любой вид связи между двумя (или несколькими) участниками* (один участник является представлением или репликой другого участника; один участник был «произведен» из других; источники создавались независимо, но отражают одну и ту же физическую систему и т.п.). Пространства данных могут *вкладываться* одно в другое, *перекрываться*, поэтому в пространстве данных должны содержаться *правила разграничения доступа*.

Отличительными *свойствами систем пространств данных* (DSSP) являются следующие:

DSSP должны работать с данными и приложениями в разнообразных форматах, доступных от многих систем через различные интерфейсы: от DSSP требуется поддержка всех данных пространства данных, без каких-либо исключений (как это бывает при использовании СУБД).

Хотя DSSP обеспечивает средства интегрированного поиска, запрашивания, обновления и администрирования пространств данных, те же самые данные часто могут быть доступны для чтения и обновления через собственный интерфейс системы, непосредственно управляющей данными: в отличие от СУБД, DSSP не имеет полного контроля над своими данными.

Могут обеспечиваться разные уровни услуг по обработке запросов к DSSP, в некоторых случаях они могут возвращать приблизительные ответы (если некоторые источники становятся недоступными, DSSP может обеспечить наилучший из возможных результат на основе данных, доступных во время выполнения запроса).

DSSP должны поддерживать средства для обеспечения более тесной интеграции данных пространства, если это необходимо.

Системы объединенных мультибаз данных, как общее решение для проблем межоперационных разнородных систем данных, обеспечивают однородный доступ к данным, сохраненным в множественных БД, которые включает несколько различных моделей данных [7, 5, 14, 16]. *Система мультибаз данных* (MDBS) – это система, которая постоянно находится «невидимо» над существующими БД и файловыми системами, называемыми локальными, и представляет для пользователя иллюзию единой БД. В MDBS поддерживается единая глобальная схема БД, в которой пользователи формируют запросы и осуществляют модификации данных, а локальные системы баз данных фактически поддерживают все данные пользователя. Глобальная схема интегрирует схемы локальных БД.

MDBS транслирует глобальные запросы в запросы к соответствующей локальной БД для фактической обработки, объединяет результаты и генерируют конечный результат для пользователя. Кроме того, MDBS координирует и аварийно прекращает глобальные транзакции для локальных систем БД, которые обрабатывают их, чтобы поддержать непротиворечивость данных внутри локальных баз данных.

Модели интеграции данных

Существующие модели интеграции данных можно разделить на три категории: модель с глобальной схемой, интегрированные базы данных и подход на основе языка мультибаз данных [11].

При реализации *модели с глобальной схемой* обязательным условием интеграции выступает наличие *глобальной схемы данных* [16]. Целью создания такой модели является интеграция данных из различных источников и предоставление пользователю унифицированного представления этих данных. Такое представление является согласованным объединением данных, к которому пользователь может адресовать запросы. В системе, основанной на такой модели, одной из главных задач является установление соответствия (отображения) между множеством источников данных и глобальной схемой. *Схема источника* описывает его структуру, в которой находятся реальные данные, а *глобальная схема* обеспечивает интегрированное виртуальное представление источников. Соответствия в *отображении* устанавливают связь между элементами глобальной схемы и схем источников. Определение виртуальной системы с глобальной схемой является достаточно общим и позволяет охватить все подходы, встречаемые в литературе. Конкретные подходы отличаются характеристиками отображения и выразительной мощностью различных языков схем и запросов.

Интегрированные базы данных основаны на моделях, в которых глобальная схема отсутствует, каждый источник представляется в структурированном виде, а *связи устанавливаются непосредственно между источниками* [11]. Такая схема предоставляет локальным базам больше возможности управления разделяемой информацией, управление получается децентрализованным. Степень интеграции не обязана быть полной, как в случае с глобальной схемой, – она зависит от потребностей пользователей, и, соответственно, такая система может быть либо сильно связанной, либо слабо связанной. Архитектура системы предусматривает наличие *общей модели данных* и *внутреннего командного языка*.

Подход *языков мультибаз данных* (МБД) предназначен для систем, которые не используют предопределенных схем интеграции. Вместо глобальной схемы на пространстве БД определяется *общее пространство имен*. Язык МБД предназначен для предоставления языковых конструкций, позволяющих выполнять запросы, которые охватывают одновременно несколько БД. Такой язык имеет возможности, которых нет в обычном языке. С его помощью пользователи определяют источники данных, способ интеграции, передачи и представления данных (пример – MRDSM с языком MSQL) [11].

Недостатками подхода с глобальной схемой являются необходимость предопределения схемы данных, статическая структура, отсутствие унифицированного представления источников данных. Недостатками подхода интегрированных баз данных являются большая автономность локальных систем, приводящая к децентрализации управления, более сложная архитектура систем управления. Недостатком подхода языков мультибаз данных является то, что задача интеграции фактически перекладывается на плечи пользователей системы, которые обеспечиваются соответствующими инструментами. Большинство систем интеграции ориентированы на *выборку информации* из разнородных источников, они не предоставляют средств манипулирования данными.

Для реализации системы управления данными, обеспечивающей интеграцию гетерогенных источников данных, был выбран подход интегрированных баз данных, поскольку он:

обеспечивает высокую гибкость системы, что позволяет динамически менять ее конфигурацию;

сохраняет достаточную степень независимости источников данных;

позволяет сформировать единое целостное представление пространства данных, абстрагируя от физических и концептуальных характеристик гетерогенной среды.

Система интеграции гетерогенных источников данных

Система управления данными должна поддерживать такие возможности, как:

многоуровневая интеграция разнородных источников данных в единое информационное пространство (т.е. интеграция не только на уровне самих источников данных, но и на уровне их элементов);

выполнение операций определения, выборки данных и манипулирования данными;

поддержка ограничений целостности на любом уровне интеграции источников данных;

поддержка произвольных связей между источниками данных на любом уровне интеграции;

представление физической схемы данных в виде логической схемы предметной области.

Далее приводится формальное описание системы управления данным, модели интеграции данных, рассматриваются свойства модели, а также архитектура системы.

Модель управления данными системы интеграции

Виртуальная система управления данными V – это тройка $<S,Q,P>$, где S – схема представления источников данных, Q – множество схем сущностей для работы с данными, P – множество отображений схемы представления источников данных на схемы сущностей:

$$F = \{p : S \to M \mid M \in Q\}.$$

Система интегрирует разнородные источники данных в единую систему, с общей схемой для реализации единого механизма управления данными, независимого от моделей данных отдельных источников.

В основе *модели данных – рекурсивное представление источников данных*. Данный способ представления позволяет моделировать произвольно сложные источники данных, сводя их к все более простым элементам. Каждый *источник данных* (на самом верхнем уровне вложенности существует один источник – сама схема) представляется в виде набора пар <*атрибут*, *источник данных*>, который разделяет данный источник на составные источники. Данные текущего уровня помещаются в атрибуты пар. Самый нижний уровень вложенности соответствует *неделимому* (*атомарному*) элементу данных. Промежуточные уровни источников данных представляют собой *метаданные*. Очевидно, что при таком подходе к моделированию обработка данных и метаданных осуществляется единообразно. Каждому источнику данных ставится в соответствие *набор связей с другими источниками* (любого уровня), *набор ограничений целостности* и *набор допустимых операций*.

В рекурсивном определении *источники данных* (ИД) обозначим $s_{l,i}^k$:

$$S = s_{0,1}^0 \text{ – источник данных самого верхнего уровня;}$$

$$s_{l,i}^k = \begin{cases} < \{< a_j, s_{i,q}^{k+1} > \left| j \in [1,J], J \in N\}, L_i^k, C_i^k(s_{l,i}^k), O_i^k >, \; k \in [0, K(i)) \right. \\ < \varnothing, L_i^k, C_i^k(s_{l,i}^k), O_i^k >, \; k = K(i) \end{cases},$$

$$i, l, q \in [1, I], I \in N, K : [1, I] \to N \cup 0$$

Источник помечается тремя индексами: l – номер родительского источника данных; i, q – номера источников данных в рекурсивном определении; k – абсолютный уровень рекурсивной вложенности; $K(i)$ – функция, возвращающая максимальный абсолютный уровень рекурсивной вложенности для данного ИД; a_j – атрибуты элементов данных (предназначены для хранения данных ИД).

Множество связей данного источника данных обозначим L_i^k, $L_i^k = \{< \{s_{l,i}^k\}, t | t \in T, n >\}$, где T – множество атрибутов связей, n – название связи; индекс i играет роль уникального ключа (у всех ИД в схеме данный индекс уникален).

Множество предикатов, задающих *ограничения целостности* на множестве источников данных, обозначается $C_i^k(S)$. Множество S может быть множеством источников данных следующего уровня вложенности, либо является вектором $\langle S_1, ..., S_n \rangle, n \in N$, где S_i – множество источников данных уровня вложенности k+1 для i-го источника данных, располагающегося на k-м уровне. Ограничения целостности для источника данных задаются в виде предикатной формулы, определенной на множестве источников данных следующего уровня вложенности либо на пространстве наборов источников данных, являющихся источниками данных следующего уровня вложенности для произвольно взятых источников данных схемы.

Источник данных $s_{l,i}^k$ находится в *согласованном состоянии*, если все предикаты из множества предикатов, задающих ограничения целостности для данного источника данных и всех источников данных, находящихся на более высоких уровнях вложенности, принимают истинное значение. Таким образом, $s_{l,i}^k$ – в согласованном состоянии, если $\forall s : s \in S_i^k \bigcup \{s_{l,i}^k\}$ выполняется условие $\forall p : p \in C(s), p(s) = true$, где s – произвольный источник данных, $p(s)$ – некоторый предикат, $C(s)$ – множество предикатов, задающих ограничения целостности для источника данных s. Источник данных находится в согласованном состоянии, если в согласованном состоянии находятся все ИД на следующем уровне рекурсивной вложенности, и он сам удовлетворяет собственным ограничениям ценности.

Рассмотрим задание различных видов ограничений целостности для источника данных *s*. Пусть *x* – переменная, обозначающая источник данных следующего уровня вложенности. Тогда:

Ограничения на диапазон допустимых значений представляются как $P(x) = a < A(x) < b,\ a, b \in M$, где A(x) – атрибут источника данных x, M – множество, в котором принимают значения атрибуты ИД.

Ссылочные ограничения: $P(x, y) = \forall x \exists y (x \to y)$.

Агрегированные ограничения: $P(x) = max(A(x)) > a, a \in M$, где A(x) – атрибут ИД x, M – некоторое множество, в котором принимают значения атрибуты источника данных.

Множество операций для данного источника данных обозначим O_i^k :

$$O_i^k = \begin{Bmatrix} \{p : S_i^{k+1} \to S_i^{k+1'}\}, k < K(i) \\ \{p : s_i^k \to s_i^{k'}\}, k = K(i) \end{Bmatrix},$$

где S_i^{k+1} – множество источников данных уровня вложенности *k+1* для *i*-го источника данных, располагающегося на *k*-м уровне.

Все *операции источников данных* могут быть разделены на четыре типа: выборка; обновление; удаление; добавление. Операции добавления и удаления некоторого источника данных определяются на множестве источников данных следующего уровня вложенности. Каждая операция может включать в себя произвольное количество операций, выполняющихся как транзакция.

Пусть *x* – добавляемый источник данных. Операция *добавления* для источника данных $s_{l,i}^k$ есть

$$f : S_i^{k+1} \to S'^{k+1}_i, k < K(i), S'^{k+1}_i = S_i^{k+1} \cup \{x\}.$$

Пусть *x* – удаляемый источник данных. Операция *удаления* для источника данных $s_{l,i}^k$ есть

$$f : S_i^{k+1} \to S'^{k+1}_i, k < K(i), S'^{k+1}_i = S_i^{k+1} \setminus \{x\}.$$

Операция *обновления* для источника данных $s_{l,i}^k$ есть

$$f : s_{l,i}^k \to s_{l,i}^{k}{}', S_i^{k+1} = S'^{k+1}_i.$$

В процессе обновления источника данных могут изменяться значения атрибутов, ограничения целостности, набор связей, набор операций. При этом изменение состава нижележащих источников данных выполняется посредством операций добавления и удаления.

Операции добавления и удаления источников данных, находящихся в согласованном состоянии, сохраняют согласованным состояние источника данных, для которого эти операции выполнялись.

В системе поддерживаются *связи* следующих типов: структурные и ассоциативные.

Структурные связи являются неявными и обусловлены рекурсивной структурой схемы данных. Декомпозиция источника данных на ряд более мелких обусловливает появление структурной связи между родительским и дочерними источниками (родительские отношения здесь рассматриваются с точки зрения вложенности источников). Структурные связи образуют *связи между разными уровнями вложенности в рамках одного фрагмента иерархии источников*, соединяют источники разных уровней рекурсивной вложенности.

Ассоциативные связи позволяют связать источники данных, которые не связаны непосредственными структурными связями. Ассоциативные связи соединяют источники одинаковых или разных уровней вложенности разных фрагментов иерархии. Если рассматривать структурные связи как вертикальные, то ассоциативные можно рассматривать как горизонтальные.

Каждая связь имеет атрибут, который характеризует действия, направленные на поддержание ссылочной целостности во время выполнения операций удаления над связанными источниками данных. Атрибут может принимать одно из следующих значений:

c – каскадное удаление: при удалении источника данных s, на который ссылается источник данных t, источник данных t также будет удален;

r – запрет удаления: если при удалении источника данных s существует источник данных t, ссылающийся на источник данных s, то операция удаления не будет разрешена; чтобы ее осуществить, необходимо либо сначала удалить источник t, либо перенаправить связь от него на другой источник;

d – удаление связи: при удалении источника данных s удаляется связь, ссылающаяся на него от некоторого источника t; при этом источник данных t продолжает существовать независимо.

Схемой сущностей называется логическое представление предметной области в виде сущностей и связей между ними, позволяющее работать с ИД схемы *S* в виде, независимом от конкретного источника данных. Каждая схема ориентирована на работу с некоторым подмножеством ИД. Определим схему сущности как $M = <E, L>$. Здесь *E* – множество сущностей, *L* – множество связей между этими сущностями ($L = <e1, e2, t>$, $e1 \in E$, $e2 \in E$, $t \in T$, где *T* – множество типов связей сущностей; ei – сущность, представляемая парой $ei = <A, O>$, где *A* – множество атрибутов, а *O* – множество операций сущности). Ключевым для сущностной схемы является отображение *p,* которое устанавливает соответствие между подмножеством ИД и схемой сущности $p : S \to E$, где *S* – некоторое подмножество схемы ИД. Для одной схемы источников данных может существовать несколько схем сущностей для решения различных задач предметной области.

Декомпозиция источника данных – это процесс последовательного структурного разбиения ИД, в ходе которого формируется рекурсивная структура представления схемы данных; при этом процесс разбиения осуществляется до тех пор, пока данные не станут далее неделимыми. Декомпозиция некоторого ИД, представленного в модели данных, в которой существуют только структурные и ассоциативные связи, может быть не единственной. Способ декомпозиции некоторого источника данных, представленного в модели данных, в которой существуют только структурные и ассоциативные связи, не влияет на максимальную глубину рекурсивной вложенности и на общее количество ИД в схеме.

Восстановление схемы данных – процесс получения схемы некоторой модели данных из схемы, полученной в результате декомпозиции схемы источников данных.

Две схемы ИД являются *эквивалентными*, если по ним восстанавливаются идентичные схемы в некоторой модели данных. Различные способы декомпозиции приводят к эквивалентным схемам ИД.

Рассмотрим возможность декомпозиции схем реляционных, иерархических и сетевых моделей данных. Как было сказано выше, декомпозиция той или иной схемы может быть не единственной.

Утверждение 1. Любая реляционная схема может быть отображена в схему источников данных.

Утверждение 2. Любая иерархическая схема может быть отображена в схему источников данных.

Утверждение 3. Любая сетевая схема может быть отображена в схему источников данных.

Данные утверждения доказываются существованием алгоритмов декомпозиции, построенных авторами.

Будем считать эквивалентными схемы реляционной, иерархической и сетевой моделей, если совпадает количество типов записей и отношений, наборы их атрибутов, за исключением первичного и внешнего ключей, а внешние ключи реляционной схемы соответствуют групповым отношениям иерархической и сетевой модели.

Утверждение 4. Любой способ декомпозиции эквивалентных схем данных иерархической, реляционной и сетевой моделей приводит к эквивалентным схемам источников данных.

Архитектура системы

Архитектура системы интеграции гетерогенных источников данных является многослойной, в ней можно выделить три уровня (рис. 1):

– уровень данных (I);
– уровень унифицированного представления данных (II);
– уровень логического представления данных (III).

На нижнем уровне системы *располагаются источники данных* и *драйверы* доступа к ним. Источники данных могут представлять собой реляционные, сетевые, иерархические, объектно-ориентированные, объектно-реляционные базы данных, электронные таблицы, XML-файлы, потоки данных, неструктурированные источники. Для доступа к источникам данных могут использоваться драйверы ODBC, провайдеры OLE DB, ADO.NET, либо иные специфические драйверы, которые поддерживают интерфейс выполнения манипуляций над источником данных, которые от него требуются в данной системе.

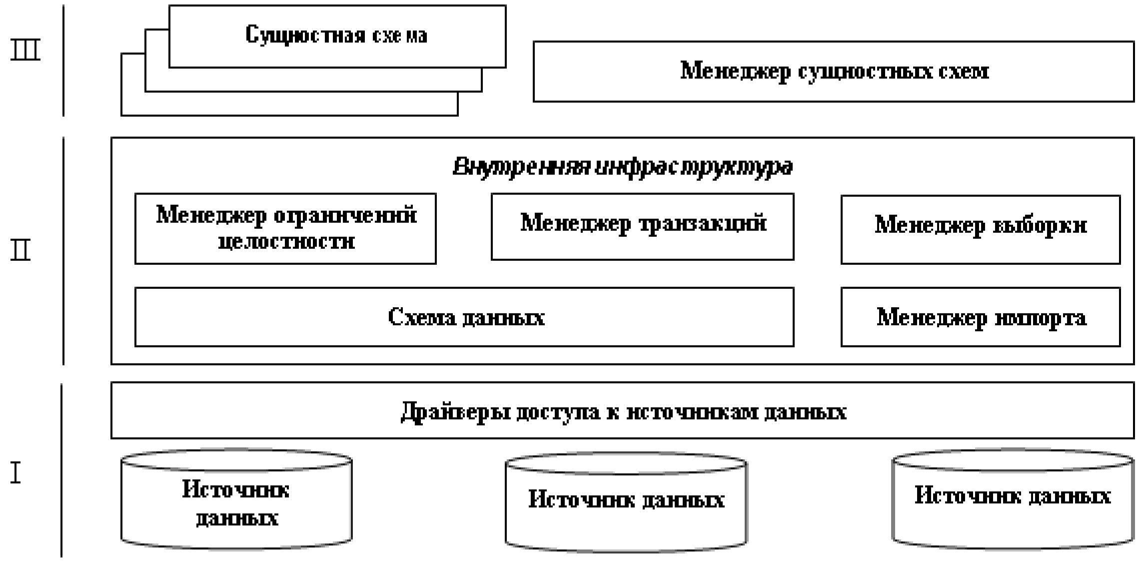


Рис. 1. Архитектура системы интеграции данных

На уровне *внутренней инфраструктуры* реализуется физическое представление схемы интеграции ИД, средства по поддержанию схемы в согласованном состоянии и выполнения базовых операций. Кроме этого, на данном уровне находится компонент, отвечающий за добавление в схему новых ИД.

Схема данных представляет собой внутреннюю рекурсивную структуру моделей источников данных в соответствии с описанной выше математической моделью.

Менеджер транзакций осуществляет выполнение глобальных транзакций в системе. Перед выполнением транзакции в объекте сохраняется информация о состоянии источников данных, которые будет охватывать транзакция, до ее начала. Это необходимо для осуществления процедуры отката в случае, если будет выполнено условие отката транзакции. На глобальном уровне транзакции выполняются последовательно и не могут пересекаться.

Менеджер ограничений целостности контролирует выполнение ограничений целостности во время выполнения операций глобальных транзакций, а также периодически с некоторым интервалом времени. Выполнение *глобальных транзакций* может состоять из произвольных операций манипулирования источниками данных, что может привести к рассогласованию данных в силу нарушения установленных ограничений целостности. Данный компонент осуществляет проверку после выполнения каждой базовой операции над источником данных на наличие нарушений ограничений целостности. Если нарушение было зафиксировано, то осуществляется откат операции и, по возможности, всей транзакции. Периодическая проверка выполнения ограничений целостности нужна на случай, если в источниках данных произошли локальные транзакции под управлением самих источников, что привело к нарушению ограничений целостности, установленных в данной схеме. В случае обнаружения такого нарушения пользователю (администратору) сообщается об этом и работа с этим ИД становится невозможной до тех пор, как согласованность данных на глобальном уровне не будет восстановлена.

Менеджер выборки данных отвечает за соединение источников данных и представление данных в виде, отвечающем потребностям пользователя. На вход компонент получает запросы, сформулированные в терминах сущностной схемы, а на выходе выдает результат выборки данных из ИД.

Менеджер импорта осуществляет импорт в систему схем данных определенного типа. В основе работы компонента лежат алгоритмы декомпозиции источников данных. Расширение спектра поддерживаемых системой моделей данных требует добавления средств интерпретации данных моделей в этот компонент.

Внешнее представление – это логическое представление интегрированной схемы, отражающее сущности и взаимосвязи конкретных предметных областей.

Схема сущностей обеспечивает работу с системой в терминах предметной области. Компонент обеспечивает высокоуровневое представление источников данных, позволяющее решать задачи предметной области, а не тратить ресурсы на интерпретацию схемы интеграции. Построение подобной схемы сущностей, по сути, означает определение отображения схемы источников данных на предметно-ориентированную схему сущностей.

Менеджер схем сущностей позволяет создавать и редактировать новые схемы сущностей и устанавливать отображение между источниками данных и сущностями.

Заключение

В ходе проведенных исследований разработана модель представления источников данных, которая позволяет осуществлять многоуровневую интеграцию гетерогенных источников данных в единое информационное пространство, обеспечивает поддержку ограничений целостности на любом уровне интеграции источников данных; поддержку структурных и ассоциативных связей между источниками данных на любом уровне интеграции.

Проведенный анализ существующих систем интеграции данных позволяет оценить преимущества предложенного подхода: возможность интеграции данных на любом уровне декомпозиции; возможность динамического изменения схемы данных; расширяемость системы за счет новых моделей данных; универсальное представление связей, ограничений целостности и операций над данными.

Продолжение работы планируется осуществлять в следующих направлениях: разработка средств разрешения проблемы семантической гетерогенности данных, механизмов поддержки распределенных транзакций в гетерогенной среде, механизмов отображения физической схемы источников данных в предметно-ориентированные схемы, алгоритмов декомпозиции и интеграции других моделей данных (в дополнение к построенным для реляционной, иерархической и сетевой моделей) и др.

Благодарности

The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine (www.aduis.com.ua).

Библиографический список

1. Агравал Р. и др. Клермонтский отчет об исследованиях в области баз данных [Электронный источник]. [Режим доступа: свободный: http://citforum.ru/database/articles/claremont_report/].
2. Кашников А.В. Средства поддержки переносимости информационных систем, основанных на метаданных. Тезисы докладов XV Международной студенческой школы-семинара – М.: МИЭМ, 2007. С.412-414.
3. Кашников А.В. Средства поддержки переносимости информационных систем, основанных на использовании метаданных // Технологии Microsoft в теории практике программирования. Новосибирск, 2007. С. 59-61.

4. Кашников А.В., Лядова Л.Н. Реализация механизма поддержания ссылочной целостности в CASE-системе METAS // Актуальные проблемы математики, механики, информатики: Материалы международной научно-методической конференции, посвященной 90-летию высшего математического образования на Урале. Перм. гос. ун-т, Пермь, 2006. С.175-177.
5. Разнородные (гетерогенные) информационные системы и системы мультибаз данных [Электронный источник]. [Режим доступа: свободный: http://www.interin.ru/page.php?id=170&pg=&print=1].
6. Blunschi L., Dittrich J.-P., Girard O.R., Kirakos S.K., Salles M.A.V. A Dataspace Odyssey: The iMeMex Personal Dataspace Management System // Conference on Innovative Data Systems Research, 2007.
7. Breitbart Y., Garcia-Molina H., Silberschatz A. Overview of multidatabase transaction management // VLDBJ, 1(2):181, 1992.
8. Carey M.J., Haas L.M., Schwarz P.M., Arya M., Cody W.F., Fagin R., Flickner M., Luniewski A.W., Niblack W., Petkovic D., Thomas J., Williams J.H., Wimmers E.L. Towards Heterogeneous Multimedia Information Systems: The Garlic Approach // Proceedings of the Fifth International Workshop on Research Issues in Data Engineering (RIDE): Distributed Object Management, 1995.
9. Chawathe S., Garcia-Molina H., Hammer J., Ireland K., Papakonstantinou Y., Unman J., Widom J. The TSIMMIS project: Integration of heterogeneous information sources // Proceedings of the 100th Anniversary Meeting. Information Processing Society of Japan, Tokyo, Japan, October 1994. P. 7-18.
10. Dittrich J.-P., Salles M.A.V. iDM: A Unified and Versatile Data Model for Personal Dataspace Management // VLDB, 2006.
11. Elmagarmid A., Rusinkiewicz M., Sheth A. Management of Heterogeneous and Autonomous Database Systems // Morgan Kaufmann Publishers, San Francisco, California, USA, 1998.
12. Franklin M., Halevy A., Maier D. From Databases to Dataspaces: A New Abstraction for Information Management // SIGMOD Record, 34(4): 2005. P. 27-33.
13. Hull R. Managing semantic heterogeneity in databases: a theoretical prospective // Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems, May 11-15, 1997, Tucson, Arizona, United States. P. 51-61.
14. Kent W. The Breakdown of the Information Model in Multi-Database Systems // SIGMOD, Record 20(4), 1991.
15. Lenzerini M. Data Integration: A Theoretical Perspective // PODS, 2002.
16. Litwin W. From database systems to multidatabase systems: Why and how // British National Conference on Databases, Cambridge Press, 1988.
17. Reddy M.P., Prasad B.E., Reddy P.G. A Methodology for Integration of Heterogeneous Databases // IEEE Transactions on Knowledge and Data Engineering, Vol. 6, No. 6, 1994. P. 920-933.
18. Sheth A.P., Rusinkiewicz M., Karabatis G. Using Polytransactions to Manage Interdependent Data // Database Transaction Models for Advanced Applications. 1992. P. 555-581,
19. Stonebraker M., Madden S., Abadi D., Harizopoulos S., Hachem N., Helland P. The End of an Architectural Era (It's Time for a Complete Rewrite). Proceedings of VLDB, Vienna, Austria, 2007.
20. Thomas G., et al. Heterogeneous distributed database systems for production use // ACM Computing Surveys, 22, 1990. P. 237-266.

Сведения об авторах

Алексей Кашников *– НИУ «Пермский государственный университет», ассистент кафедры математического обеспечения вычислительных систем; Россия, г. Пермь, 614990, ул. Букирева, д. 15; e-mail: Kashnikov@psu.ru.*
Major Fields of Scientific Research: управление данными, анализ данных, моделирование.

Людмила Лядова *– Пермский филиал Национального исследовательского университета «Высшая школа экономики», доцент кафедры информационных технологий в бизнесе; Россия, г. Пермь, 614070, ул. Студенческая, д. 38; e-mail: LyadovaLN@hse.perm.ru.*
Major Fields of Scientific Research: метамоделирование; технология DSM; CASE-средства; языковые инструментарии; предметно-ориентированные языки, DSL.

ЭКСПЕРИМЕНТАЛЬНОЕ ИЗУЧЕНИЕ ЦЕЛОСТНОСТИ ЗНАНИЙ СТУДЕНТОВ

Евгений А. Еремин

Abstract: *В работе описывается экспериментальное исследование целостности знаний, которые были получены студентами после усвоения учебного курса. Под целостностью понимается наличие связей между базовыми понятиями и терминами изученного курса, которые студент осознает и фиксирует. Предполагается, что чем больше таких связей видит студент, тем глубже он усвоил содержание курса. Мотивацией исследования данной проблемы явился поиск количественных критериев для оценки знаний студентов, в том числе и в процессе компьютерного обучения.*

Метод экспериментальной проверки целостности системы понятий был разработан и опробован автором в 2008 году (его описание было опубликовано в трудах конференции MeL-09). Компьютер помогает в реализации этого метода как при проведении проверки знаний, так и при обработке полученных результатов после ее завершения. В данной работе описаны более поздние результаты, которые были получены по предложенной методике в 2009-2010 годах. В ходе исследования проанализированы ответы более 50 студентов двух разных специальностей, которые изучали курс «Архитектура ЭВМ». Проверка знаний производилась дважды: до и после изучения курса; полученные данные сравнивались между собой.

По результатам экспериментов построены серии диаграмм для каждой из групп студентов, принимавших участие в исследовании; различные показатели сопоставлены друг с другом. Анализ результатов показал, что существуют студенты, которые лучше, чем остальные, видят связи между понятиями, причем, как выяснилось, это совсем не обязательно наиболее успешные студенты. На итоговом обобщающем графике удалось выделить четыре области, характеризующиеся разной степенью успешности освоения курса, в каждой из которых полученные в эксперименте параметры проявляют определенные особенности. В частности, довольно неожиданным явился установленный факт надежной фиксации роста целостности знаний у самых слабых студентов.

Keywords: *целостность, понятие, структура знаний, оценка, образование, курс..*

ACM Classification Keywords: *K.3.1 Computer Uses in Education; I.2.6 Learning – Knowledge acquisition.*

Введение

При проверке знаний студентов очень важно иметь четкую систему критериев, по которым преподаватель выставляет оценку. В результате многовекового опыта обучения сложились некоторые способы оценки усвоения знаний, например, в ходе беседы по изученному материалу. Тем не менее, попытки перенести эти способы в компьютерное обучение сталкиваются с существенными трудностями, поскольку экспертный опыт преподавателей довольно плохо поддается формализации.

Наиболее распространенным в данный момент методом компьютерной проверки знаний является тестирование. Типовой тест позволяет довольно быстро оценить знания тех или иных фактов, отдельных терминов или определений из изучаемого курса, но все же уступает по глубине проверки человеку-преподавателю. Конечно, тщательно разработанный тест из нескольких сотен вопросов может претендовать на объективность оценки знаний курса в целом, но проблема в том, что прохождение такого теста слишком трудоемко (опытному учителю для того, чтобы поставить оценку, требуется значительно меньше вопросов).

Можно предположить, что одним из главных преимуществ классического опроса преподавателем по сравнению с компьютерным тестом служит то обстоятельство, что человек способен по небольшому набору ответов на вопросы увидеть степень систематичности знаний экзаменуемого в целом и тем самым адекватно оценить общую картину усвоения всего курса. Справедливость предположения о важности взаимосвязей между отдельными известными студенту фактами при оценке знаний автору и хотелось проверить.

В работе экспериментально изучается целостность системы базовых понятий у студентов как один из аспектов успешности усвоения материала образовательного курса. Под **целостностью системы знаний студента** здесь и далее мы будем понимать взаимосвязь между теми понятиями и терминами, которые он усвоил в ходе изучения курса и способен продемонстрировать в ходе проверки. Подчеркнем, что нас будут в первую очередь интересовать связи между понятиями, которые осознает студент, а не логика преподавания взаимосвязанного материала курса (последняя проблема заслуживает самостоятельного изучения). В идеале все базовые термины курса должны образовывать в голове студента некоторую единую картину. Напротив, при поверхностном и некачественном изучении курса, знания студента представляют собой набор несвязанных между собой фактов.

Постановка проблемы

Рассмотрим несколько упрощенный пример, поясняющий идею исследования целостности знаний по курсу. Пусть преподаватель опрашивает трех студентов A, B и C по изученному материалу. Допустим, например, что студент A в ходе ответов на вопросы продемонстрировал знание трех следующих фактов:

- компьютер обрабатывает данные;
- для кодирования чисел используется двоичная система;
- программа хранится в памяти компьютера.

Отчетливо видно, что названные студентом факты непосредственно не связаны между собой. Предположим теперь, что ответы каждого из студентов B и C отличаются от ответа студента A всего одним фактом. Например, студент B мог отметить следующие положения:

- компьютер обрабатывает данные;
- числа – это разновидность данных;
- для кодирования чисел используется двоичная система.

А ответы студента C могли быть такими:

- компьютер обрабатывает данные;
- данные обрабатываются по программе;
- программа хранится в памяти компьютера.

Не надо обладать большим педагогическим опытом, чтобы сказать, что ответы студентов B и C лучше (хотя все три наших гипотетических студента назвали одинаковое число фактов!), поскольку они образуют некоторый *взаимосвязанный* набор фактов. Причем интересно отметить, что студенты B и C продемонстрировали свои знания в разных областях курса: студент B говорил о кодировании компьютерных данных, а студент C – об их обработке.

Таким образом, одним из показателей качественного усвоения знаний является тот факт, что студент видит и демонстрирует при контроле взаимосвязь между базовыми терминами и понятиями курса.

Сформулируем теперь наиболее важные цели нашего исследования.

- Экспериментально изучить степень связи между понятиями (целостность знаний) у студентов по изученному курсу.
- Выработать некоторые характеристики целостности знаний и опробовать их.

- Проверить, увеличиваются ли выбранные характеристики после изучения курса.
- Проанализировать, насколько целостность знаний связана с успешностью усвоения курса.

В идеале в процессе экспериментального изучения целостности знаний хотелось найти такую численную характеристику X, чтобы для студентов из рассмотренного выше примера было $X_A<X_B$ и $X_A<X_C$. В дальнейшем характеристики, подобные X, могли бы быть использованы при автоматической оценке знаний студентов.

Участники эксперимента

В ходе исследований в течение нескольких лет проверялись знания студентов физического факультета Пермского государственного педагогического университета (Россия). Для получения более достоверных результатов эксперименты проводилась на двух разных специальностях, по которым на факультете ведется подготовка: (1) учитель физики и информатики и (2) специалист по информационным технологиям (последняя предполагает, что выпускники будут обслуживать компьютерные классы в системе образования). Все эти студенты изучают курс «Архитектура ЭВМ», по которому и проверялась целостность системы базовых понятий. По первой специальности курс стоит на третьем году обучения, а по второй – на втором. Пробные эксперименты были проведены в 2008 году, а затем в 2009 и 2010 по отлаженной методике единообразно проверялись 4 группы студентов.

Данные о группах, где проводились исследования, и их численности приведены в табл. 1.

Таблица 1

группа	год	год обучения	специальность	количество студентов
G0	2008	3	физика и информатика	12
G1	2009	3	физика и информатика	12
G2	2009	2	информационные технологии	8
G3	2010	3	физика и информатика	17
G4	2010	2	информационные технологии	9
итого				58

Для удобства последующего изложения каждой из групп студентов присвоено формальное обозначение от G0 до G4. Строка с группой G0 выделена в таблице цветом, чтобы подчеркнуть, что результаты для нее содержали отдельные недоработки: из-за «не совсем чистого» проведения эксперимента они считаются предварительными.

В таблице также указано количество студентов, принимавших участие в исследовании структуры знаний. Оно не очень велико, поскольку набор по данным специальностям в последнее время сильно упал. Кроме того, были отдельные студенты, которые по уважительным причинам пропустили те дни, когда проводились экспериментальные проверки знаний, включая повторные. Результаты этих студентов оказались неполными, что, к сожалению, не позволило использовать их в исследованиях. Подчеркнем, что принимались все возможные меры, чтобы привлечь имеющихся студентов к участию в проверке знаний.

Все эксперименты по контролю знаний и вся обработка их результатов проводилась автором самостоятельно. С одной стороны, это гарантировало максимально возможную одинаковость условий проведения исследований. Но с другой – работа даже с не очень большим количеством студентов потребовала значительных усилий и времени.

База терминов для проверки знаний

Согласно замыслу эксперимента, в ходе проверки знаний студенту предъявляется полный список базовых понятий курса, а тот должен продемонстрировать, какие связи между этими понятиями он видит. Таким образом, при подготовке к проведению исследований экспериментатор должен отобрать те наиболее важные понятия и термины, которые, по его мнению, должен знать студент после усвоения курса.

Как уже говорилось выше, в качестве области исследования знаний студентов был выбран курс «Архитектура ЭВМ» (что во многом обусловлено личными предпочтениями и интересами автора). Структура базовых понятий для этого курса была подробно проанализирована ранее, и результаты этого анализа были изложены в публикации [Еремин, 2007]. Именно она и была взята для проведения экспериментов.

Опуская второстепенные детали формирования списка базовых понятий рассматриваемого курса, укажем лишь наиболее важные из них. Список состоит из 123 понятий, которые тесно связаны между собой и образуют единую иерархию терминов. В него были отобраны как наиболее общие понятия – *компьютер*, *программная* и *аппаратная часть*, *теоретические основы*, так и более конкретные термины, раскрывающие их, например, *операционная система*, *процессор*, *память*, *прямой доступ к памяти*, *принцип иерархии*, *байт* и многие другие. В список также вошли термины, реализующие межпредметные связи, например, с микроэлектроникой, логикой и системами счисления. С другой стороны, в перечень терминов сознательно не были включены названия конкретных операционных систем, внешних устройств и их производителей, а также другая подобная информация, которая является менее существенной с точки зрения изучения главных закономерностей курса. Используя стандартную терминологию, принятую в объектно-ориентированном программировании, можно сказать, что рассматривались классы понятий, но не их экземпляры.

Таким образом, был выделен достаточно широкий перечень базовых понятий, каждое из которых, с точки зрения преподавателя, должен знать и понимать грамотный студент. Перечень получился весьма объемным, так что впоследствии при проведении эксперимента оказалось, что реальные студенты активно использовали в своих ответах немногим более половины предложенного списка. Предоставление студентам такого избыточного списка имело целью сделать возможной проверку знаний у *любых* студентов, в том числе и тех, кто хорошо знаком с архитектурой компьютера или отдельными ее областями. Анализ хода эксперимента показал, что, возможно, слишком длинный перечень терминов «перегружал» студентов: они в ходе проверки уставали и демонстрировали не все свои знания. Так или иначе, но во всех описываемых экспериментах использовался именно полный список понятий из более чем ста терминов.

Для составленного списка были также предварительно проанализированы связи между отобранными терминами. Нетривиальный результат состоял в том, что удалось обойтись весьма ограниченным набором связей. В него вошли стандартные отношения между понятиями, например, *часть/целое* или *класс/подкласс*, а также некоторые специфические для курса связи вроде *основание* (*принцип иерархии* или *принцип адресации – основание – память*, *программный счетчик – основание – основной алгоритм* исполнения инструкции) и *соединение*. Полная таблица связей с конкретными примерами для каждой приведена в публикации [Еремин, 2007]: она состоит всего из 11 базовых ассоциаций.

Первоначально предполагалось, что студенты сумеют самостоятельно разобраться в таком небольшом количестве связей. Опыт тестирования, тем не менее, уже для предварительной группы G0 показал обратное: студенты часто путали даже классические *часть/целое* и *класс/подкласс*, не говоря уже об остальных видах связей. Поскольку с точки зрения цели эксперимента – общей оценки связности системы понятий, конкретные разновидности связей не так важны, было принято решение при оценке правильности ответов пренебречь ошибками в этой части задания и просто фиксировать факт наличия связи. Данное упрощение методики заметно облегчало процесс обработки и анализа результатов.

Описанная выше система терминов и была положена в основу экспериментальной процедуры проверки целостности знаний студентов по курсу архитектуры компьютера.

Описание эксперимента

В ходе эксперимента студенты *дважды* выполняли задание: в начале изучения курса и после его завершения (входной и выходной контроль). Предполагалось, что сопоставление результатов для каждого из студентов позволит индивидуально оценить степень успешности изучения курса. К сожалению, не все студенты выполняли повторное задание так же активно, как при входном тестировании, что портило результаты сравнения.

Цель эксперимента (проверка целостности знаний) не сообщалась, чтобы не вызвать искусственного повышения результатов. Студентам лишь рекомендовалось постараться выполнить задание максимально полно и хорошо, не задумываясь при этом, как и по каким параметрам будет оцениваться их успешность. Впоследствии специальная экспериментальная проверка подтвердила важность данного принципа.

Время выполнения задания не ограничивалось: каждый студент заканчивал работу индивидуально, когда считал нужным (как правило, на выполнение задания каждый студент тратил от 30 минут до часа). По-видимому, такая практика внесла некоторую неопределенность в проведение эксперимента: хотя большинство студентов старались указать как можно больше связей между понятиями курса, были и такие, кто стремился побыстрее освободиться, тем более, что выходное тестирование было последним заданием перед получением зачета по курсу. Чтобы хоть как-то уменьшить указанный негативный эффект, в последних экспериментах (группы G3 и G4) студентов предупреждали о необходимости более добросовестного выполнения второго задания. Им говорилось о том, что за время изучения курса желательно «не поглупеть».

Самый первый эксперимент, проведенный в группе G0, показал, что отдельные студенты откровенно жульничали и сдавали вместо своих результатов файлы более сильных студентов. Это стало одной из существенных трудностей проведения последующих экспериментов, хотя студентов всегда предупреждали перед экспериментом о необходимости самостоятельного выполнения задания. В 2009 году в группах G1 и G2 все файлы сравнивались между собой (сначала по длине, а потом и по содержащемуся в них тексту). Уличенным в обмане студентам группы G1 пришлось повторно выполнять работу. В небольшой группе G2, которая к тому же оказалась более добросовестной, подобных случаев зарегистрировано не было. Чтобы уменьшить усилия по контролю за достоверностью результатов и усложнить жизнь нечестным студентам, в 2010 году (группы G3 и G4) была добавлена запись в файл фамилии студента, а сам текст с результатами подвергался шифрованию. В итоге, не зная, где именно в файле хранится фамилия, студенты не могли ее заменить.

Таким образом, принимались все возможные меры для того, чтобы результаты исследования целостности знаний стали максимально объективными.

При объяснении задания студентам также выдавался текстовой файл с разъяснениями о типах связей между понятиями и их примерами. Предполагалось, что этого будет достаточно, чтобы при ответе правильно указать типы связей, Но, как уже говорилось выше, эти ожидания не оправдались и при обработке результатов типы связей игнорировались.

Кроме того, студентам предлагалось воспользоваться вспомогательным файлом с полным списком понятий, где разрешалось делать пометки и записи – тем самым участники эксперимента неявно подталкивались к систематичному выполнению задания. Судя по наблюдениям, этой возможностью никто не воспользовался.

Итак, после проведения подробного инструктажа и знакомства с программой для фиксации связей между понятиями (ее описание дано в [Еремин, 2009]), студенты приступали к выполнению задания, указывая правильные, по их мнению, связи между понятиями курса.

Результаты, сохраненные в виде файла, просматривались экспериментатором. Связи, которые были явно указаны ошибочно, из файла удалялись. Вся дальнейшая обработка файлов велась уже в автоматическом режиме.

Просмотр файлов с ответами показал, что большая часть студентов (за редким исключением) составляла связи между парами понятий весьма бессистемно. Например, в типичном студенческом файле выделялась связь между понятиями *система счисления* и *двоичная система*, но, тем не менее, отсутствовали связи этого понятия с *восьмеричной* и *десятичной системами*. На данный недостаток, как и на неумение студентов правильно определять типы связей, стоит обратить внимание всем преподавателям.

Обработка результатов эксперимента

Исходным материалом для компьютерной обработки результатов эксперимента были текстовые файлы, содержащие расшифрованные и проверенные преподавателем данные. Каждая строка такого файла содержала информацию об одной связи между понятиями и имела вид

термин 1 # тип связи # термин 2

Например, итоговый файл тестирования некоторого студента мог содержать следующий фрагмент:

процессор # целое/часть # АЛУ

процессор # целое/часть # УУ

АЛУ # целое/часть # регистр

УУ # целое/часть # регистр

Программа обработки была способна объединять связанные понятия в группы. В приведенном примере в общую группу попадают термины *процессор*, *АЛУ*, *УУ*, *регистр*. Разумеется, расположение пар терминов в файле и их порядок не играют роли.

Важно отметить, что перед анализом результаты упорядочивались по определенному рейтингу успешности освоения курса. Было принятое допущение, что *успешность определяется как срок сдачи всех заданий*: хорошие студенты, как правило, сдают задание быстро, а слабые тратят на эту же самую работу много больше времени. Сопоставление данного предположения с субъективными наблюдениями за работой студентов подтверждает, что в подавляющем большинстве случаев оно оправдывается.

Обсуждение результатов

Перейдем к обсуждению результатов эксперимента. Начнем с **первичных** характеристик, которые могут быть легко получены непосредственно из студенческих файлов с ответами. Наиболее наглядным параметром с точки зрения изучения целостности знаний является общее количество связей между понятиями курса, которые сумели вспомнить студенты. Значения этой величины для групп G1-G4 приведены в левой половине рис. 1. Каждая пара столбцов на приведенных диаграммах соответствует двум проверкам знаний: левый столбец – это количество связей L_1, указанное студентом до изучения курса, а правый – после (L_2). Студенты имеют условные номера, указанные вдоль оси абсцисс. Напомним, что согласно принятому принципу, те, кто выполнил все задания раньше (сильные студенты), имеют меньшие номера, и их столбцы расположены ближе к началу координат.

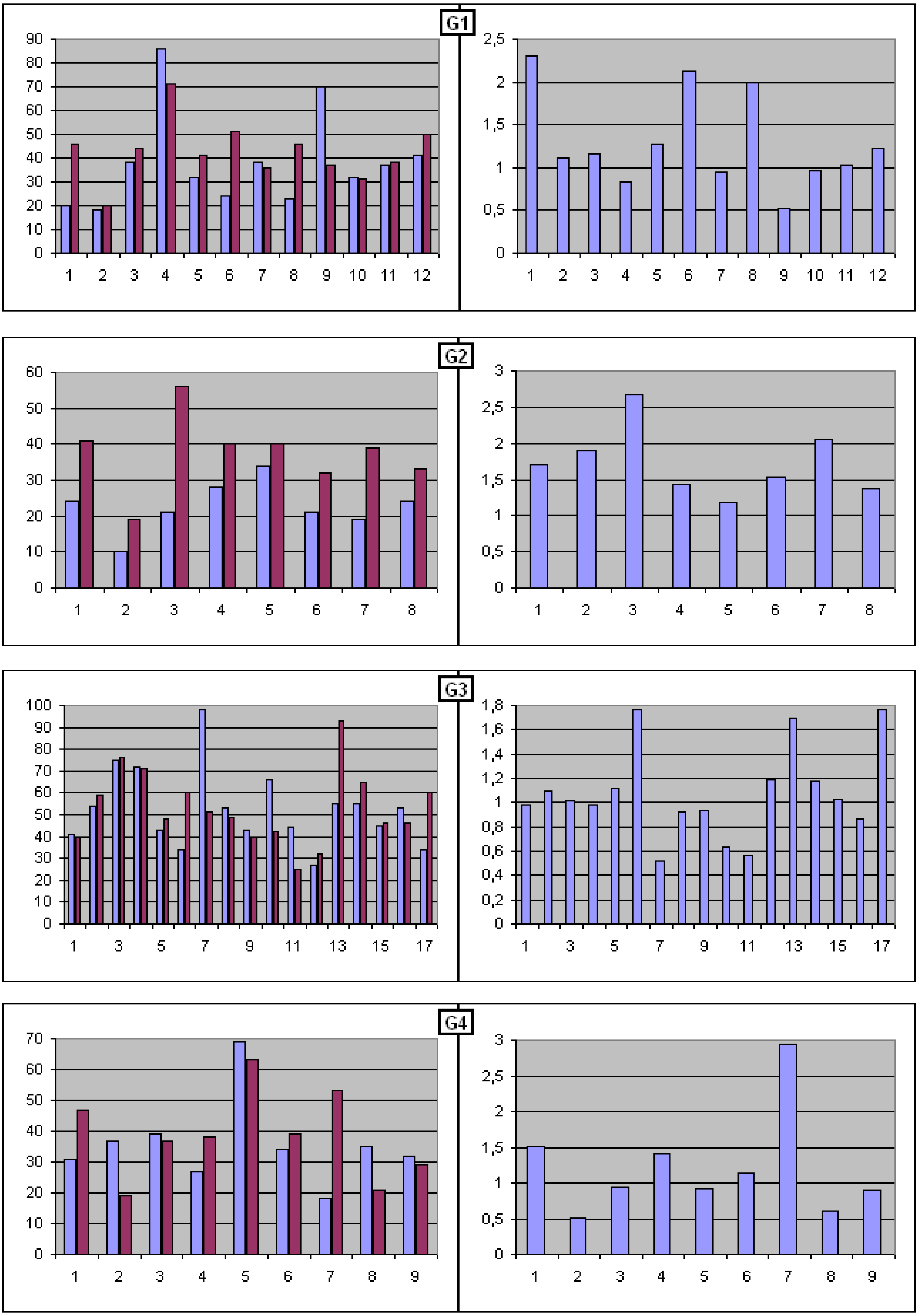


Рис. 1

Из диаграмм видно, что количество указанных студентами связей в основном попадает в диапазон от 20 до примерно 50. Лишь небольшая часть студентов имеет лучшие показатели. Причем интересно, что результаты для обеих групп специальности физика и информатика (G1 и G3) явно выше, чем для специальности информационные технологии (группы G2 и G4). По-видимому, студенты-физики лучше знают термины, связанные с электроникой и устройством компьютера. К тому же они изучают курс архитектуры на год позднее (на третьем, а не на втором курсе).

Как видно из приведенного рисунка, имеются отдельные студенты, которые умеют видеть связи заметно лучше, чем остальные. Например, студентка номер 13 из группы G3 сумела назвать после курса почти вдвое больше связей между терминами, хотя и до его изучения она фиксировала их более 50! Также можно выделить студентов номер 6 и 8 группы G1, студентку номер 3 группы G2, студенток 6 и 17 из группы G3 и студента номер 7 группы G4. Все они заметно превзошли остальных по количеству правильно указанных связей. Заметно выделяется своими результатами также студент номер 5 группы G4 (более 60 связей). Просмотр его файла показывает, что он выделяет связи очень методично: сначала указывает все связи для одного понятия, затем для другого и т.д. Несмотря на очевидную логичность такого подхода, большинство студентов группируют понятия более хаотическим образом. Подчеркнем, что все перечисленные выше студенты отнюдь не являются лидерами в группах, так что можно сделать вывод о существовании некоторой способности к анализу связей у отдельных людей. Лучшие студенты (под номером 1) из всех групп за исключением G3 также показали хорошее улучшение рассматриваемого параметра.

Поскольку наибольший интерес представляет не столько само количество связей для каждого из студентов, сколько его изменение в результате усвоения курса, в правой части рисунка построены соответствующие диаграммы для отношений количества связей после и до изучения курса: $K_L = L_2/L_1$. Теоретически количество связей, указанное при входном тестировании, не должно уменьшаться. Следовательно, значения отложенных на правых графиках отношений должны обязательно превышать единицу. К сожалению, на практике это далеко не всегда так. Разумеется, значение коэффициентов приращения «чуть меньше», чем единица ($K_L \approx 0,9$-$0,99$) можно объяснить естественной погрешностью эксперимента: студент, целостность знаний которого мало изменилась, в силу случайных факторов мог вспомнить немного меньше связей при второй проверке, чем при первой. Но это не объясняет всех «аномалий».

Рассмотрим, например, очевидный случай студента номер 9 группы G1. В ходе первой проверки он указал 70 связей между понятиями, а в ходе второй – 37 (уменьшение почти вдвое!) Ясно, что это не погрешность, а нечто другое. Внимательно посмотрев, кто еще из студентов при первом тестировании показал высокий результат, обнаружим *только одного* такого студента – с номером 4. И оказывается, что он в ходе первого эксперимента сидел рядом(!) со студентом номер 9. Таким образом, мы получаем вполне рациональное объяснение неправдоподобному результату. Конечно, наличие таких грубых отклонений не делает чести экспериментатору, который должен был следить за самостоятельностью работы студентов более тщательно. Но не стоит забывать, что многие студенты третьего курса прекрасно умеют консультироваться незаметно для преподавателя.

Отметим, что наиболее неудачными были результаты в группе G4; субъективные наблюдения за работой студентов подтверждают, что к выходной проверке эта группа отнеслась наименее добросовестно.

Попробуем теперь на базе имеющихся в нашем распоряжении первичных параметров построить более сложные **вторичные** параметры, которые получаются на основе некоторой обработки первичных. Как показало изучение предварительных результатов группы G0 [Еремин, 2009], наиболее удачный вторичный параметр получается путем следующей обработки данных. Связанные понятия объединяются в группы, как это уже было описано выше. Очевидно, что для идеального образовательного курса *все* *понятия должны войти в одну группу*, т.е. все они должны быть связанными. Реальная картина существенно менее целостна: понятия формируют несколько групп: в лучшем случае их число составляет 3-5, а в худшем – превышает 20. При этом многие группы состоят всего из 2-3 понятий, что можно смело интерпретировать как отдельные не связанные с общей картиной факты.

Если поделить общее число терминов T на число групп G, то полученное отношение будет иметь смысл среднего размера группы терминов для данного студента. Эта величина, как показывает анализ результатов, также может служить хорошей характеристикой целостности знаний студента.

Значения величин T_1/G_1 и T_2/G_2 для всех групп G1-G4 приведена на рис. 2 (номера студентов те же самые, что и на предыдущем рисунке). Графики устроены аналогичным образом: слева приведены пары значений среднего размера групп до и после изучения курса, а справа – отношения значений каждой пары, которые будем обозначать $K_{T/G}$. Основная часть выводов относительно конкретных студентов, обсуждавшихся выше, подтверждается и здесь, причем даже более отчетливо. Особенно хочется отметить, что в новых переменных становятся еще ярче видны слабые результаты группы G4 в целом и резкое превосходство в ней уже упоминавшегося выше студента номер 5. Тем не менее, есть и некоторые различия между первичными и вторичными характеристиками. Так «в новых координатах» не столь впечатляюще выглядят достижения студентки номер 3 из группы G2, а также, за исключением группы G2, снижены вторичные показатели у лидеров (студентов номер 1). Получающуюся разницу можно объяснить тем, что первичный показатель K_L, в отличие от $K_{T/G}$, никак не учитывает, насколько названные студентом связи относятся к одной и той же группе терминов. В самом деле, одинаковое значение K_L могут иметь и тот студент, который назвал множество изолированных фактов, и тот, который сумел связать абсолютно все термины в единое целое. Вторичный же параметр $K_{T/G}$ «почувствует» эту разницу, что делает его более подходящим для целей нашего эксперимента.

Анализируя рис. 2, можно заметить еще один интересный результат. Практически все слабые студенты, которые имеют большие значения номеров, обеспечили прирост измеряемой величины. Эта закономерность была отмечена еще в предварительных результатах группы G0 [Еремин, 2009]. Новые эксперименты ее подтвердили. Даже в «наиболее проблемной» группе G4 студенты с максимальными номерами 8 и 9 имеют коэффициент прироста выше, чем у остальных (он близок к единице, хотя и слегка меньше ее). Таким образом, слабые студенты явно демонстрируют в эксперименте улучшение целостности знаний.

Усредненные по группам значения некоторых первичных и вторичных показателей сведены в таблицу 2. В ней приведены: средние количества терминов T, связей L и групп терминов G, а также средние для отношений T/G и L/T. Как обычно, индексом 1 помечены результаты входной проверки, а индексом 2 – выходной. Приведенные показатели позволяют судить об уровне предварительной подготовки и результатах изучения курса всей группой в целом.

Таблица 2

группа	T_1	T_2	L_1	L_2	G_1	G_2	T_1/G_1	T_2/G_2	L_1/T_1	L_2/T_2
G1	48	51	38	43	13	12	4,14	4,50	0,78	0,82
G2	32	46	23	38	10	11	3,37	4,23	0,71	0,81
G3	62	60	52	53	13	13	5,10	5,35	0,83	0,87
G4	42	45	36	38	8	11	6,05	5,54	0,85	0,84

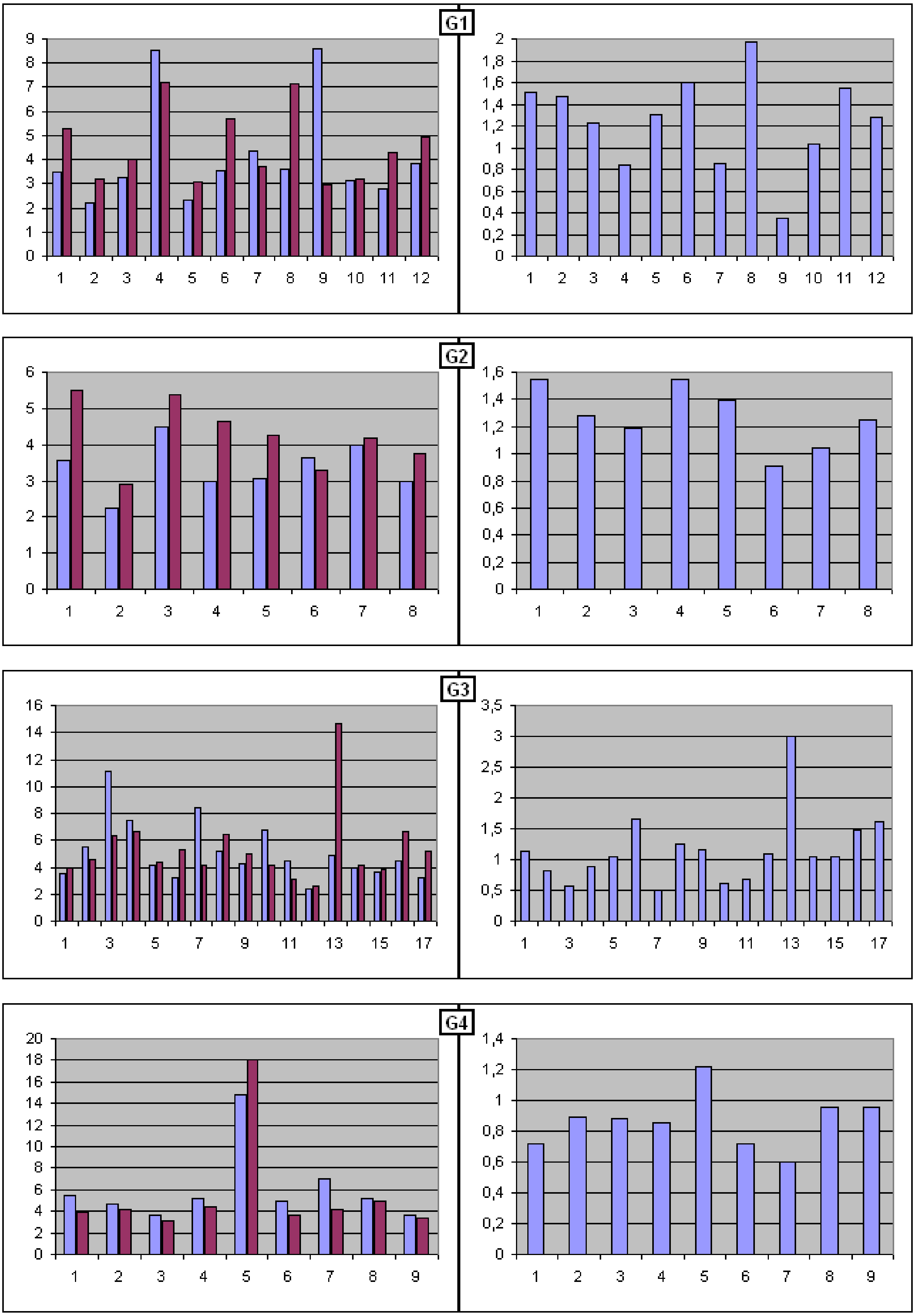


Рис. 2

Наконец, еще один наглядный способ описания, предложенный автором в предварительной публикации [Еремин, 2009], – это специальные диаграммы целостности знаний (см. рис. 3). Такая «пятнистая» диаграмма построена следующим образом. Каждая точка в столбце диаграммы – это одно понятие. Столбец разделен на полосы, которые являются графическим представлением сформировавшихся у студента групп понятий. Для улучшения наглядности группы имеют чередующуюся белую и серую окраску, а нижняя, самая большая, которая образует своеобразное базовое ядро из понятий курса, – черную. Нетрудно также заметить, что группы во всех столбцах упорядочены по величине, так что самая большая группа всегда находится внизу, а самые маленькие, состоящие из 2-3 понятий – в верхней части столбика.

Таким образом, приведенная на рисунке комплексная визуальная форма представления информации сочетает в себе следующие сведения:

- высота столбика пропорциональна общему числу усвоенных понятий, которые имеют хотя бы одну связь с другими понятиями;
- количество разноцветных областей характеризует степень «разрозненности» (фрагментарности) системы понятий;
- размер групп связанных понятий (на рисунке характеризуется площадью соответствующих областей) также свидетельствует о целостности представлений в той или иной части курса;
- размер нижней, самой большой, области (на рис. 3 она выделена черным) отражает объем базового блока понятий курса.

В идеале диаграмма должна представлять собой однородный (состоящий из одной группы) столбик черного цвета, причем его высота для списка, применявшегося в наших экспериментах, должна соответствовать более чем 120 терминам, т.е. почти вдвое выше, чем это получалось на практике.

Используя приведенные на рис. 3 диаграммы, можно получить более наглядную картину успешности освоения курса студентами. Если рис. 2 позволяет сравнить целостность знаний только по одному усредненному показателю, то «пятнистая» диаграмма гораздо лучше отражает распределение понятий по группам. Она служит своеобразной «картой знаний», глядя на которую преподаватель может сделать определенные выводы об усвоении своего курса.

Рассмотрим в качестве характерного примера студентку номер 5 группы G2. Ее количественный показатель – общее число сгруппированных понятий практически не изменился. (Если вернуться к рис. 1, то и увеличение общего количества связей у данной студентки тоже невелико.) Зато отчетливо видно, что структура знаний значительно улучшилась: выросла базовая группа понятий (выделенная на рисунке черным цветом), а количество мелких групп (символизирующих изолированные факты), напротив, уменьшилось. Заметим, что показатель $K_{T/G} \approx 1{,}4$ (см. рис. 2) значителен, т.е. улучшение целостности знаний есть несмотря на то, что количество связей изменилось мало.

С точки зрения преподавателя, наиболее удачными следует признать результаты, продемонстрированные в группе G2. В ней все студенты не только показали прирост количества связанных понятий, но и улучшили структуру своих знаний – диаграммы выходной проверки выглядят более целостно. Обращает на себя внимание тот факт, что G2 является самой маленькой группой (всего 8 человек); возможно, это обстоятельство привело к более успешному контролю самостоятельности выполнения заданий.

В то же время подробность диаграмм на рис. 3 одновременно оборачивается их недостатком. Причина в том, что хотя наглядность диаграммы и удобна для преподавателя-человека, она трудно применима в автоматизированных обучающих системах. В самом деле, на что в первую очередь следует обращать внимание: на размер максимальной группы терминов или на общее количество групп? А, может быть, за основу лучше принять меньшее количество «мелких» изолированных групп? На эти вопросы трудно дать однозначный ответ.

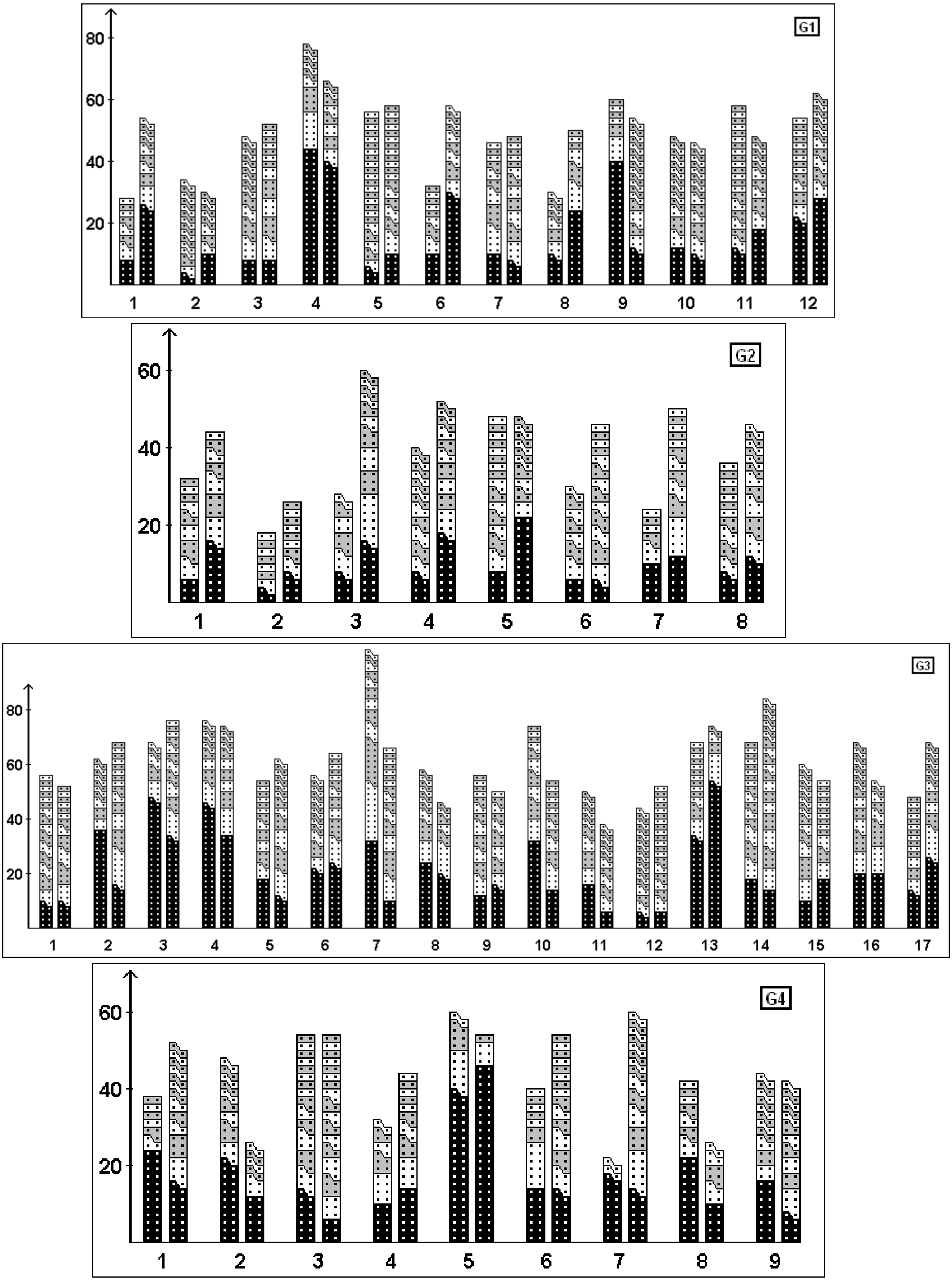

G1
80
60
40
20
1 2 3 4 5 6 7 8 9 10 11 12
G2
60
40
20
1 2 3 4 5 6 7 8
G3
80
60
40
20
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
G4
60
40
20
1 2 3 4 5 6 7 8 9

Рис. 3

Из результатов группы G4 снова выделяется систематичностью своих знаний студент номер 5, о котором уже шла речь выше.

При построении рассматриваемых диаграмм был также проведен дополнительный эксперимент, целью которого было проверить влияние на его результаты знания студентами цели проверки. Для этого четырем самым недобросовестным студентам группы G1, уличенным в сдаче чужих файлов было предложено не просто повторить тестирование, но и через некоторое время выполнить задание еще один (третий!) раз. Но в последнем случае они уже знали, что такое целостность знаний, в чем состоит цель тестирования и как оцениваются его результаты. Студентам было предложено обратить внимание на то, чтобы связать между собой все известные им понятия. В результате размер базовой группы понятий возрос в 2-4 раза, но единой группы все равно не получилось! Конечно, это были не лучшие студенты, но результаты все равно свидетельствуют о том, что целостность знаний не приходит по желанию.

Как показали приведенные выше результаты, достижения студентов, связанные с целостностью знаний, не всегда однозначны и имеют индивидуальные особенности. В частности, ранее уже было отмечено, что рост количественных показателей надежно регистрируется у слабых студентов. Попробуем проанализировать полученные данные с целью выяснить, нет ли других связей между экспериментально полученными параметрами и успешностью освоения курса. Обратимся к значениям прироста среднего размера группы терминов $K_{T/G}$, которые приведены в правой части рис. 2.

К сожалению, группы G1-G4 имеют разное число студентов и данные необходимо предварительно нормировать. Для этого можно предложить следующий несложный прием. Введем некоторую формальную величину X, которая будет описывать нормированный номер студента. Самому первому студенту поставим в соответствие значение X=0, а самому последнему – X=1. Для остальных студентов значения X нетрудно вычислить по формулам

$$\Delta X = 1/(N-1), \quad X_n = (n-1)\Delta X, \quad \text{где } n = 1 \ldots N$$

Здесь n – это номер студента, а N – общее количество студентов в группе. Вспомогательная величина ΔX имеет смысл расстояния между соседними точками для студентов данной группы; например, для группы G4 число студентов N=9 и ΔX = 0,125. Для других групп значения ΔX будут другими, причем чем больше N, тем меньше ΔX (иными словами, точки вдоль оси абсцисс будут стоять плотнее).

Нормированные таким способом точки для всех четырех групп G1-G4 сведем в единый график $K_{T/G}(X)$, изображенный на рис. 4.

Его нижняя область ($K_{T/G}$ < 1), закрашенная серым цветом, соответствует «некорректным» результатам (студент ухудшил свой показатель в результате изучения курса). Напомним также, что точки с малыми значениями X вблизи нуля соответствуют студентам, быстро и успешно освоившим курс, а большие X, напротив, отражают результаты слабых студентов. Глядя на полученный график, можно довольно уверенно выделить на нем 4 зоны; обозначим их Z1-Z4.

Зона Z1 соответствует самым сильным студентам. В ней имеется довольно значительный разброс значений $K_{T/G}$ от 0,5 до 1,5. Это наиболее способная часть студентов в группах, получение зачета для них не составляет особого труда. Поэтому результат здесь существенно зависят от степени заинтересованности в выполнении задания. Отдельные способные студенты настолько не стремятся показывать высокие результаты, так что даже получают $K_{T/G}$ < 1.

Зону Z2, занимающую чуть меньше половины используемого отрезка оси X, можно назвать зоной средних студентов. Это студенты, которые учатся успешно и, как правило, выполняют все данные им задания добросовестно. Если отбросить единичную точку с $K_{T/G} \approx 0,5$ как нехарактерную, то большая их часть демонстрирует прирост целостности знаний. Некоторые точки ниже единицы вполне можно отнести к погрешности педагогического эксперимента.

Следующая зона Z3 демонстрирует даже больший разброс значений $K_{T/G}$, чем Z1: значения здесь «прыгают» от $K_{T/G} < 0,5$ до 3. Это слабые и (или) недобросовестные студенты, занимающиеся нестабильно и кое-как. Результаты это подтверждают.

Наконец, зона Z4 – это самые слабые студенты, которые, как сразу бросается в глаза, все обеспечивают прирост своих (не очень больших) знаний. По-видимому, для этой группы студентов улучшать свои знания – это *единственный* способ получить зачет. В частности, их уровень знаний низок, поэтому любые попытки обмануть преподавателя сразу же бросаются в глаза, так что остается только в меру своих сил заниматься.

Таким образом, получается очень интересная картина распределения параметра $K_{T/G}$, вычисляемого из экспериментальных данных: две зоны Z1 и Z3 содержат трудно предсказуемые результаты, зато две другие (Z2 и Z4) оказываются стабильными. Данный результат, к сожалению, свидетельствует о том, что разработанная методика экспериментального определения целостности знаний не всегда дает надежные результаты.

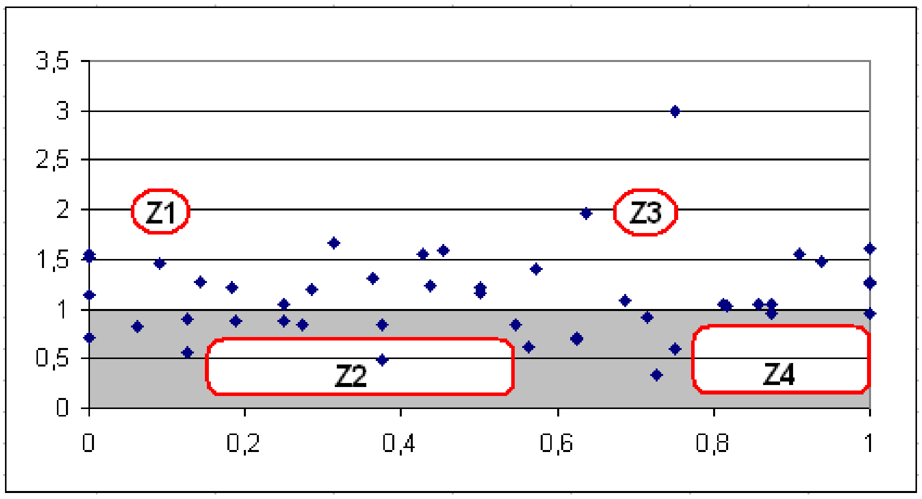


Рис. 4

Заключение

По проведенному анализу результатов эксперимента можно сделать выводы как по совершенствованию самой методики эксперимента, так и по целостности знаний у протестированных студентов.

Опыт проведения эксперимента позволяет выдвинуть следующие предложения по совершенствованию экспериментальной методики проверки связанности понятий, сформированных после изучения курса.

Поскольку, как показали наблюдения за ходом работы, выполнение дважды одного и того же задания нравится не всем студентам, можно предложить упростить проведение эксперимента: проводить проверку целостности знаний только после завершения курса, а результаты сопоставлять с некоторыми средними значениями, полученными ранее (например, воспользовавшись таблицей 2). По мнению автора, эта простая мера позволит улучшить объективность контроля, хотя и в ущерб возможности оценки индивидуального прироста знаний из-за отсутствия входного контроля.

При проведении итоговой проверки в самом конце курса некоторые студенты торопились сдать свои результаты. Поэтому, возможно, целесообразно ввести ограничение на минимальное время выполнения задания (можно, например, поставить условие, что время повторного тестирования должно быть не менее времени первичного).

В ходе проведения проверки знаний необходимо тщательно следить за самостоятельностью работы студентов и принимать все меры для предотвращения сдачи чужих файлов. Здесь хорошо зарекомендовало себя применение шифрования содержимого итоговых текстовых файлов.

Результаты показывают, что базу понятий для тестирования не стоит брать слишком большой. Есть предположение, что использованная в работе база была чрезмерно велика. В пользу этого вывода свидетельствует тот факт, что подавляющее большинство студентов использовали менее половины из предложенного списка понятий.

Все перечисленные рекомендации будут обязательно проверены в ходе дальнейших исследований целостности знаний студентов.

Перейдем теперь к выводам по полученным экспериментальным данным.

Эксперименты отчетливо показали, что почти все(!) студенты не умеют определять типы связей между понятиями. Они даже путают классические взаимосвязи *целое/часть* и *класс/подкласс*, не говоря уже о других нестандартных связях. По-видимому, преподавателям стоит обратить внимание на подобную ограниченность мышления студентов и развивать их возможности в этом направлении (разумеется, это относится к любому учебному курсу, а не только к курсу компьютерной архитектуры).

Как свидетельствуют результаты эксперимента, после изучения курса целостность знаний у разных студентов изменяется по-разному; существуют студенты, которые заметно лучше, чем остальные, видят связи между понятиями.

В связи с предыдущим выводом, становится понятным, почему не удалось обнаружить какую-то универсальную характеристику целостности знаний. Сопоставление различных характеристик позволяет предложить в качестве наиболее удачного параметра среднее количество терминов в группе. Хорошую помощь преподавателю при углубленном анализе результатов оказывают диаграммы группировки понятий, предложенные автором.

Интересным результатом является выявленное в результате экспериментов разделение всех студентов на несколько групп (зон) в зависимости от уровня подготовки. К сожалению, не во всех из них показатели целостности знаний измеряются надежно. Особо хочется подчеркнуть обнаруженные в исследовании зоны средних и очень слабых студентов, показатели в которых демонстрируют стабильное поведение.

Acknowledgement

The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine (www.aduis.com.ua).

Bibliography

[Еремин, 2007] E.A.Eremin. Using Topic Map technology in the planning of courses from the CS knowledge domain. In: Proc. Seventh Baltic Sea Conference on Computing Education Research (Koli Calling 2007). CRPIT, 88. ACS, 2007.

[Еремин, 2009] Е.А. Еремин. О компьютерной методике изучения целостности системы базовых понятий, сформировавшейся у студентов в результате освоения курса. В: Human Aspects of Artificial Intelligence. IBS "INFORMATION SCIENCE & COMPUTING", N 12, vol. 3. Institute of Information Theories and Applications FOI ITHEA, Sofia, 2009.

Author's Information

Evgeny A. Eremin *– Senior lecturer, Perm State Pedagogical University, 614990, Sibirskaya st., 24, Perm, Russia; e-mail: eremin@pspu.ac.ru*

Major Fields of Scientific Research: Knowledge representation, Computer Learning, Computer architecture

TABLE OF CONTENTS