Преобразования реляционных схем и задача "восстановления грамматики"

Ткачев Игорь Иванович,
с.н.с., доцент
Научно-учебный центр прикладной информатики НАН Украины

К задаче "восстановления грамматики" на примере реляционных схем

грамматика $G \rightarrow$ язык L_G

язык $L \rightarrow$ грамматика G_L

<u>Определение.</u>

Пусть задано n переменных $u_1, u_2, \ldots, u_n$ и n множеств $A_1, A_2, \ldots, A_n$, которые являются доменами их значений.

Подмножество их декартова произведения будем называть *n-арным отношением* на множестве переменных $u_1, u_2, \ldots, u_n$:

$R \subseteq A_1 \times A_2 \times \ldots \times A_n$; $R = \{(a_1, a_2, \ldots, a_n)\}$

При этом множество $u_1, u_2, \ldots, u_n$ называют *диапазоном отношения scope (R)*

$u_1, u_2, \ldots, u_n$
$a^1{}_1, a^1{}_2, \ldots, a^1{}_n$
. . .
$a^s{}_1, a^s{}_2, \ldots, a^s{}_n$

<u>Определение.</u>

Множество отношений (таблиц) будем называть реляционной схемой.

Задача удовлетворения ограничений (ЗУО)

Constraint Satisfaction Problem (CSP)

Постановка задачи удовлетворения ограничений (УО), дискретный случай.

ЗУО может рассматриваться как тройка (V, D,C), где

$V = \{x_1,..., x_n\}$ — множество дискретных переменных,

$D = \{D_1,..., D_n\}$ — множество доменов переменных,

$C = \{C_1,..., C_m\}$ множество ограничений.

Ограничением называется пара (R, S) где R – отношение, определенное на диапазоне S, $S \subseteq V$.

Решением ЗУО называется присвоение значений всем переменным, которое удовлетворяет всем ограничениям. *Целью* решения ЗУО может быть нахождение одного или всех решений.

Задача удовлетворения ограничений (ЗУО)

Приложения ЗУО и программирования в ограничениях:

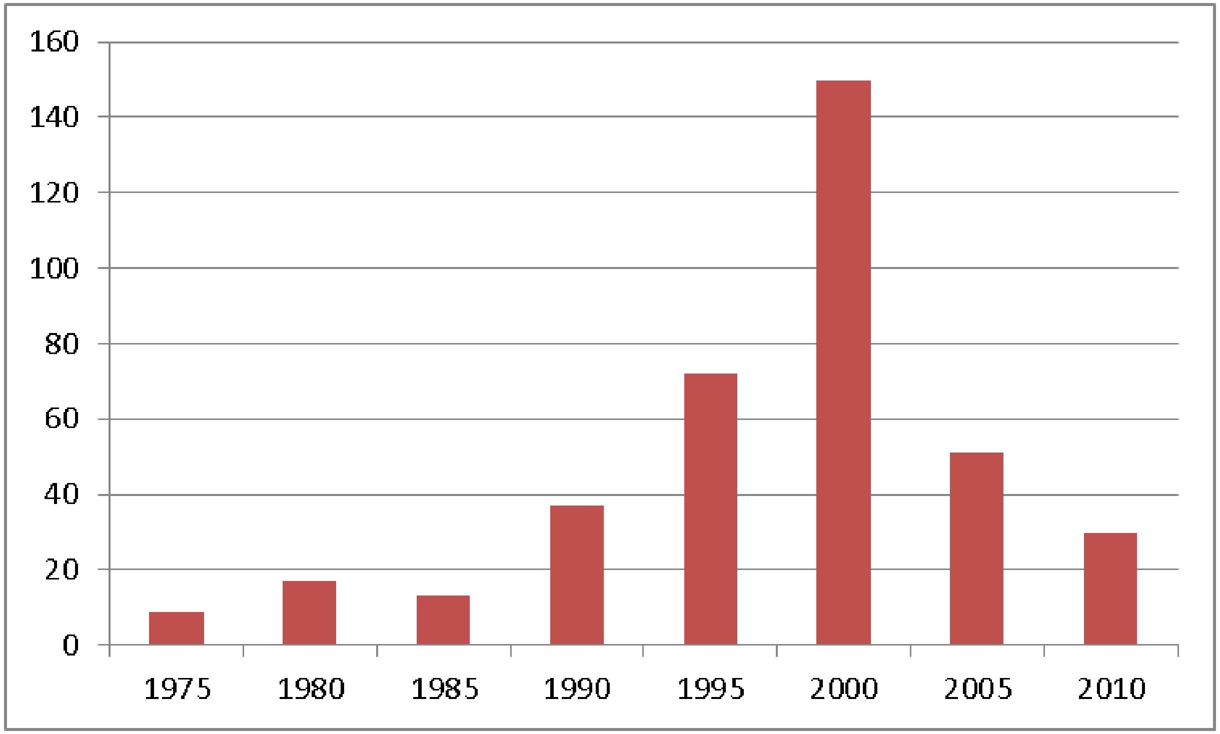
- Графовые задачи (раскраска графов, изоморфизм)
- Задачи пропозициональной выполнимости
- Логическое программирование
- Имитационное моделирование систем
- Анализ электронных схем
- Обработка естественного языка
- Проектирование экспертных систем
- Вероятностные сети
- Составление/решение кроссвордов, судоку и других занимательных задач

Пример. Составление ограничения для задачи раскраски графов

V1	V2
1	2
1	3
1	4
2	1
2	3
2	4
…	…

V2	V3
1	2
1	3
1	4
2	1
2	3
2	4
…	…

Задача удовлетворения ограничений (ЗУО)



Число научных публикаций по теме УО

Задача состоятельной разметки (СР)

Consistent Labeling Problem

Схемой разметки называется четверка <U, L, W, R>, где
U – конечное множество объектов,
L – конечное множество меток,
W – конечное множество неупорядоченных наборов объектов $(u_1, u_2, \ldots, u_n)$,
R – объединение множеств R^i, где R^i – множество наборов вида $(u_1, l_1, u_2, l_2, \ldots, u_n, l_n)$.
Набор меток $(l_1, l_2, \ldots, l_n)$ из набора $(u_1, l_1, u_2, l_2, \ldots, u_n, l_n)$ называется разметкой набора объектов $(u_1, u_2, \ldots, u_n)$.

Состоятельной относительно W и R разметкой будем называть функцию $f: U \to L$, такую, что для любого набора объектов $(u_1, u_2, \ldots, u_n) \in W$ выполняется:

$(u_1, f(u_1), \ldots, u_n, f(u_n)) \in R$.

Множество всех состоятельных разметок схемы S=<U, L, W, R> обозначим $\mathcal{L}$(W, R).

Задача СР1. ? ($\mathcal{L}(W, R) = \varnothing$)

Задача СР2. ? $f: U \to L \in \mathcal{L}(W, R)$

Задача СР3. ? $\mathcal{L}(W, R)$

СР1 $\in$ NP

Задача состоятельной разметки (СР)

$S=<U, L, W, R>$

$u_l, u_2, \ldots, u_n$	. . .	$v_l, v_2, \ldots, v_n$
$l_{11}, l_{12}, \ldots, l_{1n}$ ..... $l_{s1}, l_{s2}, \ldots, l_{sn}$	....	$k_{11}, k_{12}, \ldots, k_{1n}$ ..... $k_{r1}, k_{r2}, \ldots, k_{rn}$

<u>Замечание 1.</u> Задача СР эквивалентна задаче УО.

<u>Замечание 2.</u> В общем случае множество «диапазонов» W может быть составлено из наборов разной длины.

<u>Замечание 3.</u> Задача СР с диапазонами ограничений произвольной арности (длины) может быть сведена к задаче СР с бинарными диапазонами.

<u>Замечание 4.</u> Схема задачи СР совпадает со схемой реляционной БД.

<u>Замечание 5.</u> Задача СР представляет собой задачу поиска подмножества гомоморфизмов отношений.

Задача состоятельной разметки (СР) и задача гомоморфизмов отношений

<u>Определение</u>. Пусть заданы два n-арных отношения – $W \subseteq A^n$ и $T \subseteq B^n$. Гомоморфизмом из W в T называется такое отображение ϕ: A→B, что для любого кортежа $(a_1, a_2, \dots , a_n) \in W$ выполняется $(\phi(a_1), \phi(a_2), \dots , \phi(a_n)) \in T$.
Множество всех гомоморфизмов обозначим $\mathbf{\Gamma}(W,T)$.

Пусть заданы два n-арных отношения – $W \subseteq A^n$, A = {a} и $T \subseteq B^n$, и B = {b}.

Построим схему разметки S_{WT}=<A, B, W, R>, где

$R = \{(a_1, b_1, a_2, b_2, \dots , a_n, b_n) \mid (a_1, a_2, \dots , a_n) \in W; (b_1, b_2, \dots , b_n) \in T\}$.

W =	w_1	…	w_m
	T	…	T

Тогда $\mathcal{L}(W, R) = \mathbf{\Gamma}(W,T)$.

Таким образом, в общем случае СР представляет собой задачу установления некоторого подмножества множества гомоморфизмов двух отношений. Это следует из того факта, что переход от множества R к его подмножеству может привести к потере части состоятельных разметок.

<u>Замечание.</u> Переходя от отношений к системам отношений, т.е. реляционным схемам, получаем этот же результат, т.е. задача поиска гомоморфизмов реляционных схем ставится как задача СР, причем сама схема СР также является реляционной схемой.

Задача состоятельной разметки (СР)

S=<U, L, W, R>

$u_1, u_2, \ldots, u_n$	. . .	$v_l, v_2, \ldots, v_n$
$l_{11}, l_{12}, \ldots, l_{1n}$ $l_{s1}, l_{s2}, \ldots, l_{sn}$	……..	$k_{11}, k_{12}, \ldots, k_{1n}$ $k_{r1}, k_{r2}, \ldots, k_{rn}$

Заметим, что некоторую состоятельную разметку можно рассматривать как большой кортеж:

U =	$u_1, u_2, \ldots\ldots\ldots, u_N$
f: U→L	$f(u_1), f(u_2), \ldots\ldots\ldots, f(u_N)$

В каждом столбце исходной таблицы есть разметка, являющаяся фрагментом состоятельной разметки.

Аналогично можно изобразить все множество состоятельных разметок:

U =	$u_1, u_2, \ldots\ldots\ldots, u_N$
	$\mathcal{L}(W, R)$

Это также табличный вид задачи СР. Причем с тем же самым множеством состоятельных разметок.

S_0=<U, L, {U}, {$\mathcal{L}$(W, R)}>

Схема состоятельной разметки (СР) как порождающая конструкция

Две стороны схемы СР:

1. Схема есть запись задачи поиска соответствия между двумя структурными объектами (отношениями, реляционными схемами).
2. Схема СР может рассматриваться, как конструкция, «задающая» / «допускающая» множество состоятельных разметок. Схема исполняет роль «грамматики», а «языком» является $\mathcal{L}$(W, R).

Проверка допустимости заданной схемой S=<U, L, W, R> некоторой функции f: U→L.

Если для $\forall$ w $\in$W выполняется (w, f(w))$\in$ R, то f $\in\mathcal{L}$(W, R).

Другими словами, по каждому диапазону из W проектируем f и проверяем наличие полученного кортежа меток в соответствующем столбце.

Задача восстановления схемы разметки (СР)

Задача восстановления схемы разметки (реляционной схемы /реляционной грамматики).

Пусть U – конечное множество объектов, $|U|=N$;

L – конечное множество меток.

Задан «язык» К в виде N-арного отношения.

Требуется построить схему $S=\langle U, L, W, R\rangle$ так, что $\mathcal{L}(W, R) = K$.

U =	$u_1, u_2, \ldots , u_N$
	К

$\rightarrow$

W=	w_1	w_2	.....	w_m
	R_1	R_2		R_m

Эквивалентные преобразования схемы разметки (СР)

Определение.

Две схемы разметки S_1=<U, L, W_1, R_1> и S_2=<U, L, W_2, R_2>, определенные на одних и тех же множествах объектов и меток U, L, называются *эквивалентными*, если $\mathcal{L}(W_1, R_1) = \mathcal{L}(W_2, R_2)$.

Эквивалентное преобразование состоит в переводе схемы S_1=<U, L, W_1, R_1> в эквивалентную ей схему S_2=<U, L, W_2, R_2>.

Первое эквивалентное преобразование: вычеркивание несущественной разметки

Определение. Две разметки (l_1, l_2, ... , l_n) и (k_1, k_2, ... , k_n) диапазонов (u_1, u_2, ... , u_n) и (v_1, v_2, ... , v_n), соответственно, называются противоречивыми, если $\exists$ i, j такие, что $u_i = v_j$ и $l_i \neq k_j$.

Определение. Разметка (l_1, l_2, ... , l_n) диапазона $w \in W$ схемы S=<U, L, W, R> называется несущественной, если найдется такой другой диапазон $w' \in W$, все разметки которого ей противоречат.

Удаление из схемы S=<U, L, W, R> любой несущественной разметки означает переход к эквивалентной схеме S=<U, L, W, R'>, так как такая разметка не может быть частью какой-либо состоятельной разметки.

Последовательное удаление несущественных разметок есть классический алгоритм вычеркивания («релаксации»).

Эквивалентные преобразования схемы разметки (СР)

Второе эквивалентное преобразование: дописывание несущественной разметки

Дописывание в схему S=<U, L, W, R> любой несущественной разметки означает переход к эквивалентной схеме S=<U, L, W, R''>, так как такая разметка не может быть частью какой-либо состоятельной разметки.

Вычеркивание несущественной разметки и дописывание несущественной разметки являются парой взаимообратных преобразований.

Операции с отношениями: проекция

Второе эквивалентное преобразование: дописывание несущественной разметки

Проекция $\Pi_Y(R)$ отношения R на множество переменных $Y \subseteq$ scope (R) представляет собой отношение, состоящее из частей кортежей исходного отношения, соответствующих диапазону Y.

		Y	
M =	$u_1, \dots$	$u_j, \dots, u_k$	$\dots, u_n$
	$l_1, \dots$	$l_j, \dots, l_k$	$\dots, l_n$
R=	. . .	. . .	. . .
		$\Pi_Y(R)$	

Операцию проектирования можно обобщить на случай одновременного проектирования на систему множеств $(Y_1, Y_2, \dots, Y_m)$, $Y_i \subseteq$ scope (R), $i=1,\dots,m$.

Операции с отношениями: проекция и соединение отношений

Соединение (*join | composition*) двух отношений R_1 и R_2 дает в результате новое отношение $R_3 = R_1 * R_2$ и выполняется следующим образом:
Диапазоны исходных отношений объединяются scope (R_3) = scope $(R_1) \bigcup$ scope (R_2);
Кортежи нового отношения составляются из всех пар кортежей исходных отношений R_1 и R_2, которые согласуются по переменным из пересечения диапазонов - scope $(R_1) \bigcap$ scope (R_2).

<u>Замечание</u>.
(1) Пусть R_1 и R_2 состоят только из кортежей, которые имеют согласованный кортеж в другом отношении. Тогда
$(R_1$ и $R_2) \rightarrow R_3 = R_1 * R_2 \rightarrow \Pi_{\text{scope }(R1)}(R_3) = R_1$ и $\Pi_{\text{scope }(R2)}(R_3) = R_2$;

(2) $(R) \rightarrow \Pi_{\text{scope }(Y1)}(R)$ и $\Pi_{\text{scope }(Y2)}(R)$, $\{ Y1 \bigcup Y2 = \text{scope }(R) \} \rightarrow R' = R_1 * R_2$; $R' \neq R$.

Соединение двух «согласованных» отношений R_1 и R_2 и дальнейшее проектирование не приводит к потере исходной информации.
Проектирование отношения приводит к результату, по которому, в общем случае, нельзя восстановить исходное отношение.

Операции с отношениями: проекция и соединение отношений

Пример 1.

a b	b c
1 1	1 2
1 2	2 4
3 2	2 5

→

a b c
1 1 2
1 2 4
3 2 4
1 2 5
3 2 5

→

a b	b c
1 1	1 2
1 2	2 4
3 2	2 5
1 2	
3 2	

Это всего лишь повтор

Пример 2.

a b c
1 1 2
1 2 4
3 2 5

→

a b	b c
1 1	1 2
1 2	2 4
3 2	2 5

→

a b c
1 1 2
1 2 4
1 2 5
3 2 4
3 2 5

А это необоснованная добавка

Эквивалентные преобразования схемы разметки (СР)

Третье эквивалентное преобразование СР: соединение согласованных столбцов

Можно показать, что соединение согласованных столбцов есть эквивалентное преобразование схемы разметки.

Замечание 1. Соединение согласованных столбцов в отличие от первых двух операций меняет множество диапазонов W и множество ограничений R схемы S=<U, L, W, R>.

Замечание 2. Согласование столбцов есть не что иное, как вычеркивание несущественных разметок в выбранной паре столбцов.

Замечание 3. Последовательное применение вычеркивания несущественных разметок в выбранной паре столбцов и последующее их соединение в конце процедуры даст один столбец с набором состоятельных разметок. Фактически это один из вариантов алгоритма поиска СР. Порядок выбора пар с точки зрения конечного результата несущественен.

Эквивалентные преобразования схемы разметки (СР)

Четвертое эквивалентное преобразование СР: расшивание столбца *(decomposition)*

Для выполнения этой операции столбец должен иметь определенную структуру множества разметок.

Кортеж переменных исходного столбца представим в виде $\{w_1cw_2\}$, где w_1, c, w_2 – его подкортежи, а $\{w_1c\}$ и $\{c\ w_2\}$ – кортежи переменных пары результирующих столбцов.

$\mathbf{w_1cw_2}$
B_1
B_2
. . .
B_m

→

$\mathbf{w_1c}$	$\mathbf{cw_2}$
$B_1^{\,1}$	$B_1^{\,2}$
$B_2^{\,1}$	$B_2^{\,2}$
. . .	. . .
$B_m^{\,1}$	$B_m^{\,2}$

Каждая разметка w_1 сочетается с каждой разметкой w_2

Расшивание столбца является эквивалентным преобразованием СР. *Соединение согласованных столбцов* и *расшивание* столбца являются парой взаимообратных преобразований.

$\mathbf{w_1}$	**c**	$\mathbf{w_2}$
k_1	l_c	h_1
. . .		. . .
k_1	. . .	h_n
k_2		h_1
. . .		. . .
k_2		h_n
. . . .		. . . .
k_m		h_1
. . .	l_c	. . .
k_m		h_n

m различных разметок, каждая повторяется по n раз

n различных разметок, каждая повторяется по m раз

Эквивалентные преобразования схемы разметки (СР)

Теорема.

Четыре преобразования схемы разметки –

- вычеркивание несущественной разметки;
- дописывание несущественной разметки;
- соединение согласованных столбцов и
- расшивание столбца

образуют полную систему преобразований.

Это значит, что для любых двух эквивалентных схем

S_1=<U, L, W_1, R_1> и S_2=<U, L, W_2, R_2> существует последовательность из указанных преобразований, позволяющая перейти от одной схемы к другой.

Эквивалентные преобразования схемы разметки (СР)

Доказательство.

1). Наряду с S_1=<U, L, W_1, R_1> и S_2=<U, L, W_2, R_2> рассмотрим схему S_0=<U, L, {U}, {$\mathcal{L}$(W, R)}>.

2). Последовательно для всех столбцов схем S_1 и S_2 проводим вычеркивание и соединение согласованных столбцов вплоть до получения схемы S_0. Получаем две последовательности, из вычеркиваний несущественных разметок и соединений согласованных столбцов $S_1 \to S_0$ и $S_2 \to S_0$.

3). Используя последовательность $S_2 \to S_0$, строим в обратном направлении последовательность $S_0 \to S_2$, заменяя соединения согласованных столбцов двойственными им расшиваниями, а вычеркивания несущественных разметок – дописываниями.

4). Соединяя последовательности $S_1 \to S_0$ и $S_0 \to S_2$, получаем $S_1 \to S_2$.

Замечание.

Очевидно, что существует множество последовательностей преобразований для перехода от одной СР к другой эквивалентной СР: $S_1 \to S_2$.

Полная система эквивалентных преобразований СР и задача восстановления схемы разметки

Схема S_0 фактически является исходным условием задачи восстановления СР.
Из приведенного доказательства следует, что применяя преобразования из введенной выше полной системы эквивалентных преобразований, можно получить любую эквивалентную схему, в том числе, и ту, что нас удовлетворит по условию «компактности» (минимальности).

Требуется разработать алгоритм, а также формализовать условия «компактности» задачи восстановления СР. Остается открытым и вопрос, когда исходные условия допускают достижение «компактности». К примеру, если «язык» состоит из одного «предложения», все СР кроме исходной будут менее компактны.

<u>Замечание.</u>
Алгоритм для восстановления схемы должен на каждом шаге находить такие «несущественные» разметки для дописывания, которые в дальнейшем дадут возможность новых расшиваний. В таком аспекте эти разметки становятся весьма «существенными».

Задача восстановления схемы разметки

Если бы нам было известно множество диапазонов W_x искомой схемы S_x=<U, L, W_x, R_x>, то множество R_x мы могли бы получить (а с ним – и всю схему) посредством проектирования единого столбца исходной схемы S_0 по множеству W_x.

Фигурально выражаясь, правильно подобранная система диапазонов раскрывает/отображает «сложность» моделируемого отношения (= «языка»).

Фактически на языке теории отношений задача ***восстановления схемы разметки*** есть задача ***представления отношения системой его проекций.***

Задача восстановления схемы разметки

Выше на примере было показано, что соединение отношений не является операцией обратной к проецированию.
Из этого факта, следует, что не всякое отношение может быть задано системой проекций с заданным множеством диапазонов W.

Вопрос: если множество диапазонов W задано, то, что можно сказать о множестве состоятельных разметок («языков»), порождаемых схемами вида
S_x=<U, L, W, R_x>?
Назовем такие языки (множества СР/отношения) *W-представимыми* – их множество обозначим M_W.

1) $\varnothing \in M_W$, – наименьший элемент;
2) $L^N \in M_W$, где N = |U|– наибольший элемент;
3) (единичные кортежи длины N из элементов L) $\in M_W$;

Получаем частично упорядоченное множество по включению.
Если $A \in M_W$ и $B \in M_W$, то $A \cap B \in M_W$. И это есть операция взятия нижней грани.
Определим операцию взятия верхней грани следующим образом:
Проецируем $C = A \cup B$ по множеству диапазонов W, затем находим множество СР: C^*; $C \subseteq C^*$. Это будет верхняя грань.
Таким образом, множество всех W-представимых языков есть решетка.

Выводы

1. Рассмотренные задачи лежат в русле направления, развивающегося вокруг задачи удовлетворения ограничениям (ЗУО, англ. Constraint Satisfaction Problem). Здесь же находится задача состоятельной разметки (СР, англ. Consistent Labeling Problem).
2. В общем случае задача СР представляет собой задачу установления некоторого подмножества множества гомоморфизмов двух реляционных систем (в частности, отношений).
3. Схема разметки может рассматриваться как порождающая конструкция. Порождаемым продуктом является множество отображений $\mathcal{L}$(W, R) ={f: $U \to L$}.
4. В связи с п.3 правомерна постановка задачи о построении СР, порождающей заданное множество состоятельных разметок. На языке теории отношений это задача о представлении отношения системой его проекций.
5. Постановка задача восстановления СР требует доработки в плане уточнения условий, которым должна удовлетворять искомая СР (типа компактности или сокращения объема памяти для ее записи).
6. Введено понятие эквивалентных преобразований СР и построена полная система из четырех эквивалентных преобразований СР. (В реляционной алгебре Кодда два преобразования из этой системы отсутствуют).
7. Система введенных преобразований представляется перспективным инструментом для анализа и решения задачи восстановления СР.
8. Введено понятие *W-представимости* множества СР. Показано, что далеко не каждое множество СР может быть задано схемой с заданной системой диапазонов *W.* Показано, что множество W-представимых множеств представляет собой *решетку*.

Публикации автора по теме

1. Ткачев И.И. Эквивалентные преобразования в одном классе распознающих систем. - Кибернетика, 1981, No 1, с. 27-32.

2. Ткачев И.И. Реляционные методы распознавания: две фундаментальные постановки // "Математические методы распознавания изображений" тезисы докладов, Киев: Ин-т киб. АН Украины, 1991, 2 с.

3. Ткачев И.И. Реляционные методы распознавания образов и задача гомоморфизма алгебраических моделей // Доклады АН Украинской ССР Сер. А, 1988, №1, с.76–78.

4. Ткачов І.І. Реляційні моделі для опису структур складних об´єктів у розпізнаванні образів // Міжнародна науково-практична конференція «Наука, освіта, суспільство: актуальні питання і перспктиви розвитку», 29-30 січня 2016 р., м. Одеса. – Одеса: ГО «ІОМП», 2016. – С. 164–169. (http://novaosvita.com.ua/wp-content/uploads/2017/01/ScEdSoc-Kyiv-Dec2016_P2.pdf)

WEB: http://nucpi.nas.gov.ua/center-staff/item/21-tkachev-ihor-ivanovych.html

E-mail: tk_i@ukr.net

Литература по ЗУО

1. Шлезингер М. И. Синтаксический анализ двумерных зрительных сигналов в условиях помех.— Кибернетика, 1976, №4, с. 113—130.
2. Haralick R. M. Scene Matching Problems. – In: Digital image processing and analysis: NATO advanced study institute, 1978, p. 184–202.
3. Rosenfeld A., Hummel R. A., Zucker S. W. Scene Labeling by Relaxation Operations.– IEEE Transactions on System, Man and Cybernetics, 1976, SMC—6, N 7, p. 420—433.
4. Handbook of Constraints Programming. – Edited by F.Rossi, van Beek and T.Walsh, Elsevier B.V., 2006, 955 p.
5. Kumar V. Algorithms for constraint satisfaction problems: A survey // AI Magazine. 1992. 13(1). P. 32-44.
6. Dechter R., Frost D. Backtracking algorithms for constraint satisfaction problems. Tech.rep., University of California, Department of Information and Computer Science, 1999. (http://www.ics.uci.edu/~csp/r56-backtracking.pdf)
7. Bartak R. Theory and practice of constraint propagation // Proceedings of the Third Workshop on Constraint Programming for Decision and Control (CPDC-0I). Poland, Gliwice, 2001. P. 7-14.
8. Meseguer P. Constraint satisfaction problems: an overview // AI Communications. 1989. V. 2., P. 3-17.
9. Miguel I., Shen Q. Solution techniques for constraint satisfaction problems: advanced approaches // Artificial Intelligence Review. 2001. V. 15. P. 267-291.
10. Рассел С., Норвиг П. Искусственный интеллект: современный подход. 2-е изд. / Пер. с англ. М.: Вильямс, 2006. 1408 с. (http://rriai.org.ru/zadachi-udovletvoreniya-ogranicheniy.html)

СПАСИБО ЗА ВНИМАНИЕ!