УДК 510.62

А. Ф. ОСЫКА, канд. техн. наук, *И. Н. ВОРОНИНА*

О РАСПОЗНАВАНИИ ЭЛЕМЕНТОВ ЗНАЧЕНИЯ ТЕКСТА

СООБЩЕНИЕ 2

Машинная реализация процедур анализа запросов на естественном языке, вводимых в информационную систему, предъявляет ряд требований: обеспечить единообразие представления различных сетей переходов, используемых в данной работе для анализа фрагментов запросов, чтобы общий алгоритм состоял из возможно меньшего числа стандартных блоков; вы-

брать такую структуру данных, используемых в процедурах анализа, чтобы в случае увеличения количества сетей переходов, их расширения, изменений в условиях восприятия словоформ сетью и т. п. не требовалось менять процедурную часть системы анализа запросов; обеспечить возможность настройки системы для анализа запросов на естественном языке в различных тематических областях без изменения самих процедур анализа [1—3].

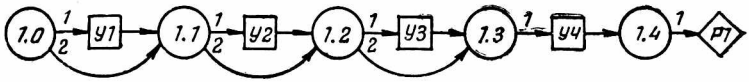

Рис. 1

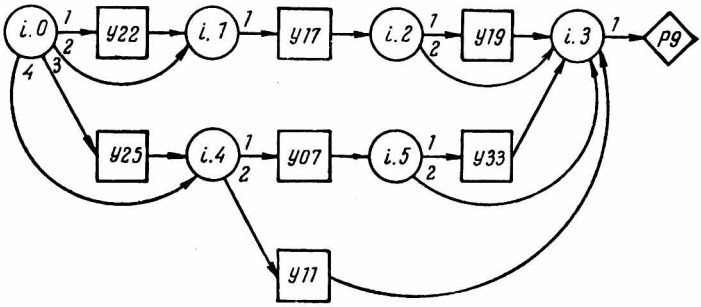

Рис. 2

Рассмотрим способ унифицированного представления сетей переходов на примере сетей с номерами 1 и *i*, изображенных на рис. 1, 2.

Подобные сети (а их в системе анализа запросов для некоторой тематической области может быть около ста) удобно представлять в виде совокупности трех таблиц: переходов, условий и результатов. Первая содержит в себе данные о всех сетях системы. Для сетей, изображенных на рис. 1, 2, она имеет вид табл. 1.

В графе 1 помещаются номера сетей, используемых в данной информационной системе. Каждая сеть представлена несколькими строками в табл. 1 — по одной на каждое проверяемое условие в этой сети. В графе 2 помещается номер условия, выполнение которого следует проверить на очередном этапе анализа запроса с помощью данной сети. Содержание проверяемых признаков раскрыто в таблице условий (табл. 2) в строке, номер которой совпадает с содержимым данной строки в графе 2 табл. 1. В графах 3, 4 помещены сведения о том, какие действия следует выполнять в случае положительного

Таблица 1

№	№ проверяемого условия	Действия по «да»		Действия по «нет»		№ записи
1	2	3	4	5	6	7
Символы						
3	3	1	3	1	3	1
001	001	У	002	У	002	1
001	002	У	003	У	003	2
001	003	У	004	У	004	3
001	004	Р	001	В		4
i	022	У	017	У	017	1
i	017	У	019	У	025	2
i	019	Р	009	Р	009	3
i	025	У	007	У	007	1
i	007	У	033	У	011	2
i	033	Р	009	Р	009	3
i	011	Р	009	В		2

результата проверки условия из графы *2*. Содержимое графы 4 в сочетании с символом «У» в графе 3 данной строки означает номер условия, к проверке которого следует перейти на следующем этапе анализа сетью. Запись в графе 4 в сочетании с символом «Р» означает номер строки в таблице результатов (табл. 3), откуда следует считать результат, получаемый на выходе данной сети. Символ «В» в графе 3 означает безрезультатное прекращение анализа этой сетью.

Графы 5, 6 содержат сведения о действиях, которые следует выполнить в случае отрицательного исхода проверки условия из графы 2 в данной строке. Структура и содержание информации в этих графах такие же, как в 3, 4.

Таблица 2

№ условия	Перенос	Место в предложении			Проверка 1	Проверка 2	Проверка 3
1	2	3	4	5	6	7	8
Символы							
3	1	1	1	2	10	10	10
001	—	1	+	05	Сем=СУБ	—	—
002	—	0	1	02	Сем=МОД	—	—
003	—	0	1	02	Сем=ИМЕТЬ	Чр=глаг	—
004	—	0	1	05	Сем=МЕСТО	—	—
022	1	2	1	02	Сем=МЕСТО	Чр=сущ	Род=муж
022	—	—	—	—	Числ=ед	Пад=вин	—

Графа 7 (табл. 1) содержит номера *j* записей, которые заносятся во вспомогательный массив ЗАП при проверке очередного условия данной сети. Если проверка условия из графы 2 (табл. 1) дала положительный результат, то во вспомогательном массиве производится запись ЗАП(*j*) = «*j*, Н(*j*), Слов(*j*), Сем(*j*)», где *j* — содержимое графы 7 (табл. 1) в строке, соответствующей проверяемому условию; Н(*j*) — номер словоформы во фразе запроса, выявленной проверяемым условием; Слов(*j*) — словоформа запроса, выявленная запросом; Сем(*j*) — семантический признак выявленной словоформы, который присвоен ей при морфологическом анализе. Если проверка очередного условия в сети дала отрицательный результат, то во вспомогательный массив заносится запись ЗАП(*j*) = «*j*, —, —, —».

Табличное представление сети переходов строится на основе графического. Но таблица содержит сведения не только о самой сети, но и о порядке работы с нею: к какому условию перейти в случае положительного или отрицательного исхода проверки предыдущего условия, а также какую запись вспомогательного массива оставить, а какую — убрать. Это позволяет упростить процедуру анализа, так как снимает вопрос о поиске путей в сети, о запоминании пройденных ветвей, о хранении полученных данных и т. п.

Таблица 3

№ результата	Результат		№ записи	Метка
	Признак	Значение		
1	2	3	4	5
	Символы			
3	5	7	1	1
001	$SIT=$	С1	—	—
002	$TOP=$	Сем (*j*)	2	—
003	$POD=$	Слов (*j*)	1	—
004	$TOT=$	Сем (*j*)	2	—
005	$NOT=$	Слов (*j*)	3	—
006	$SIT=$	С2	—	1
006	$TOP=$	Сем (*j*)	4	1

Сведения о содержании проверок, которые следует выполнить в очередном условии сети переходов, помещены в таблице переходов. Пусть в сети переходов (рис. 1) У1 = {Сем = СУБ, Мвп = 1 + 0,5}, У2 = {Сем = МОД, Мвп = 0102}, У3 = {СЕМ = =ИМЕТЬ, ЧР=глаг, Мвп=0102}, У4={СЕМ=МЕСТО, Мвп= =0105}. Эти условия, а также условия У22 = {СЕМ = МЕСТО, Мвп = 2102, Чр = сущ, Род = муж, Число = ед, Пад = вин} представлены в табл. 2.

В графе 1 (табл. 2) помещаются номера условий, которые проверяются на очередном этапе анализа сетью. Пробел в графе 2 означает, что данное условие занимает одну строку. Запись одного условия может занимать несколько строк. Тогда в каждой строке этого условия (кроме последней) в графе 2 проставляется какой-либо символ, не равный пробелу. (См., например, условие № 022).

В графах 3—5 указываются координаты места в запросе, где следует искать словоформу с необходимыми признаками. Символ «1» в графе 3 означает, что поиск словоформы следует вести от начала фразы. «0» в этой позиции указывает на то, что поиск выполняется относительно предыдущей словоформы, выявленной данной сетью, при отсутствии таковой — от начала фразы. «2» означает, что поиск следует вести относительно словоформы запроса, которой приписан номер данной сети. «1» в графе 4 предписывает вести поиск словоформы справа и слева от точки отсчета, указанной в графе 3 данной строки. «—» в графе 4 означает, что поиск ведется слева от точки отсчета, а «+» — справа. В графе 5 указывается, на сколько словоформ можно удаляться от точки отсчета при поиске требуемой словоформы.

Содержимое граф 6—8 заранее не оговаривается. В каждую из этих граф может быть записан код любого необходимого признака и через разделитель «=» — его значение. Это дает возможность при необходимости изменять содержание некоторого условия путем замены в табл. 2 соответствующей строки на одну или несколько строк уточненного условия под прежним номером.

В отдельных случаях при описании условия требуется указать качества, которыми словоформа в данной позиции не должна обладать, чтобы словосочетание было воспринято анализирующей сетью. Такое описание условия с отрицанием в терминах рассматриваемой сети переходов может быть выполнено двумя способами. При первом в графах 6—8 (табл. 2) проставляются коды признаков с отрицанием. Такие записи обрабатываются специальными процедурами. При втором способе условие, которое не должно выполняться, при переходе из соответствующего круглого узла сети проверяется первым без отрицания. Если это условие выполнено, то анализ сетью прекращается или происходит переход к условию на другой ветви сети, не включающей данный круглый узел. Второй способ предпочтительнее, так как не требует дополнительных процедур обработки записей условий в табл. 2.

Данные, получаемые в результате положительного исхода анализа каждой сетью, содержатся в табл. 3, которая представляет собой фрагмент таблицы результатов для системы обработки запросов о железнодорожных билетах.

Каждая сеть на выходе (табл. 3), как правило, дает свой результат, который чем-то отличается от результатов анализа других сетей. Поэтому в табл. 3 будет примерно столько различных номеров в графе 1, сколько имеется различных номеров в графе 1 табл. 1. В графе 2 «Признак» (табл. 3) помещены переменные типа *SIT* (ситуация), *TOP* (тип места), *POD* (пункт отправления), *TOT* (тип поезда) и т. п., значение которых существенно для «понимания» запроса. В графе 3 указано конкретное значение, которое должна принимать переменная из графы 2 в данном результате. Если в некоторой строке графы 3 проставлен код, не равный Сем(j) или Слов(j), то этот код берется в качестве значения переменной из графы 2. Если же в графе 3 стоит код Сем(j) или Слов(j), то значение переменной формируется. Для этого берется запись из вспомогательного массива, номер j которой считывается из графы 4 в данной строке. Из записи № j считывается тот реквизит, который указан в графе 3 табл. 3. Значение этого реквизита принимается в качестве значения переменной из графы 2.

В результате работы некоторых сетей получает значение только одна переменная, а на выходе остальных сетей выдаются значения нескольких переменных. Например, если сеть

настроена на сочетания типа: «Есть ли билеты...», то результат равен {*SIT*-С2}. Если же сеть распознает словосочетания типа: «Есть ли купейные билеты...», то получается результат {*SIT* = С2, *TOP* = Сем(4)}. Результат, содержащий значение только одной переменной, занимает одну строку в табл. 3. Результат, состоящий из значений нескольких переменных, записывается в несколько строк (по количеству определяемых переменных). Если запись некоторого результата продолжается в следующей строке, то в графе 5 табл. 3 проставляется любой символ, отличный от пробела (см., например, результат № 006). Считывание значений переменных из каждой строки производится по общему правилу.

Обозначим через Т4(k, l), Т1(m, n), Т2(p, q), Т3(r, s) соответственно содержимое таблиц: результатов морфологического анализа (табл. 4), переходов (табл. 1), условий (табл. 2) и результатов (табл. 3). В табл. 4 в графе Т4(k, 2) записывается словоформа запроса, в Т4(k, 3) — код семантического класса словоформы, в Т4(k, 4) — признак части речи, в Т4(k, 5) — остальная морфологическая информация о словоформе, в Т4(k, 6) — Т4(k, 10) — номера сетей переходов, которые анализируют возможные контексты данной словоформы запроса. ЗАП(j) означает запись j во вспомогательном массиве. С учетом введенных обозначений алгоритм анализа запросов в информационную систему может быть представлен следующим образом.

1. Получение таблицы Т4(k, l), $k = 1, 2, ..., K$, $l = 1, 2, ..., 10$. (Выполнение морфологического анализа запроса).

2. Считывание очередной записи Т4(k, l)$\neq$0, где k=1, 2, ..., K, l = 6, 7, 8, 9, 10. (Определение очередного номера сети из таблицы результатов морфологического анализа).

3. Выбор первой строки в Т1(m, n), где Т1(m, 1) = Т4(k, l). (Считывание первой строки таблицы переходов, относящейся к рассматриваемой сети).

4. Считывание и проверка условия Т2(p, q), в котором p = Т1(m, 2), по таблице условий.

5. Условие Т2(p, q) выполнено? Если «да», то переход к п. 6, если «нет» — к п. 11.

6. ЗАП(j) = «j, Н(j), Слов(j), Сем(j)», где j = Т1(m, 7). (Во вспомогательный массив заносится результат проверки данного условия).

7. Т1(m, 3) = Р? (Выполненное условие позволяет получить результат на выходе сети?) По «да» переход к п. 17, по «нет» — к п. 8.

8. Т1(m, 3) = В? (Выполненное условие ведет в тупиковую ветвь сети, соответствующую отрицательному условию?) Если «да», то переход к п. 18, «нет» — к п. 9.

9. Выбор строки m' в $T1(m, n)$ такой, что $T1(m', 1) = T4(k, l)$ и $T1(m', 2) = T1(m, 4)$. (Выбираем строку с номером условия, к которому следует перейти в случае положительного исхода проверки очередного условия в данной сети).

10. $m = m'$. (Дальнейшей обработке подлежит вновь выбранная строка таблицы переходов). Переход к п. 4.

11. ЗАП(j) = «j, —, —, —», где $j = T1(m, 7)$. (Во вспомогательный массив заносится фиктивная запись).

12. $T1(m, 5) = P$? (Невыполнение условия позволяет получить результат на выходе сети?) Если «да», то переход к п. 15, «нет» — к п. 13.

13. $T1(m, 5) = B$? (Невыполнение условия вызывает необходимость прекратить анализ данной сетью? По «да» — переход к п. 18, по «нет» — к п. 14.

14. Выбор строки m' в $T1(m, n)$, в которой $T1(m', 1) = T4(k, l)$ и $T1(m', 2) = T1(m, 6)$. (Считывание строки с номером условия, к которому необходимо перейти при отрицательном исходе проверки очередного условия в данной сети). Переход к п. 10.

15. $r' = T1(m, 6)$. (Следует воспользоваться строкой результата в табл. 3, номер которой взят из графы 6 таблицы переходов).

16. Получение и запись результата работы данной сети по строке $T3(r', s)$ таблицы результатов. Переход к п. 18.

17. $r' = T1(m, 4)$. (Обрабатывается строка из табл. 3, номер которой содержится в графе 4 таблицы переходов). Переход к п. 16.

18. $k = K$ и $l = 10$? (Обработаны все записи с номерами сетей в табл. 4?) Если «да» — выполняется п. 19, «нет» — переход к п. 2.

19. Отбор, упорядочивание и перекодировка переменных и их значений, существенных для ситуации запроса. Конец работы алгоритма.

В результате работы алгоритма получается набор семантических переменных и их значений, которые удобно использовать как указания на соответствующую процедуру поиска в информационной системе, а также на параметры и их значения, необходимые для работы этой процедуры.

При составлении программы для ЭВМ дальнейшей детализации требуют лишь блоки 4 и 16. Записи во вспомогательном массиве ЗАП могут быть использованы для оценки достоверности результата анализа сетью. Для этого результату анализа на выходе сети ставится в соответствие числовой коэффициент $\leqslant 1$. Он равен количеству нефиктивных записей в массиве ЗАП, деленному на количество всех записей, произведенных данной сетью. Если имеются несовместимые значения переменных, полученные в результате работы различных

сетей, то выбирается значение, у которого больше коэффициент на выходе сети.

При переходе от одной ограниченной тематической области к другой в различных информационных системах достаточно изменить словарь системы, используемый для морфологического анализа, а также содержимое табл. 1—3. Действия с данными этих таблиц не меняются.

Список литературы: 1. *Шенк Р.* Обработка концептуальной информации.— М.: Энергия, 1980.— 356 с. 2. *Диалоговые* системы в АСУ / Под ред. Поспелова Д. А.— М.: Энергия, 1983.— 216 с. 3. *Попов Э. В.* Общение с ЭВМ на естественном языке.— М.: Наука, 1982.— 360 с.

Поступила в редколлегию 25.12.85