МАТЕМАТИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМ
(Вычислительные системы)
1981 год Выпуск 89

УДК 519.681.2+519.765:519.682

ОБ ОДНОМ МЕТОДЕ СТРУКТУРИРОВАНИЯ ТЕКСТОВ

И.Н. Скопин

При решении задач преобразования текстовой информации с помощью ЭВМ часто удобно пользоваться внутренней, содержательной структурой обрабатываемых текстов. Традиционные языки, предназначенные для обработки текстов, такие как СНОБОЛ [1] и РЕФАЛ [2], представляют средства для описания преобразований в терминах поиска некоторых фрагментов текста по образцу (строке) и замены их на другие строки. Эти средства, по существу, моделируют алгорифм Маркова [3] и потому принципиально пригодны для описания практически любых текстовых преобразований. Однако их использование наталкивается на существенные трудности программирования, когда описание алгоритма преобразований проще и естественней задавать, используя структуру текста. По этой причине в РЕФАЛ, например, введены специальные средства, которые удобны при скобочном структурировании текстов. Тем не менее скобочная структура является всего лишь одной из возможных структур и не всегда достаточна для задания преобразований.

Предлагаемый ниже метод структурирования текстов позволяет определять практически любую структуру обрабатываемого материала с учетом его содержания и с целью описания алгоритмов в рамках введенной структуры. Структурирование задается с помощью набора строковых функций, вычисление которых на обрабатываемой строке порождает разбиения ее на так называемые структурные единицы. Множество всех структурных единиц представляет собой полную совокупность строковых объектов, используемых для описания обработки. Это множество частично упорядочено. Каждой структурной единице соответствует некоторый кортеж целых неотрицательных чисел (структурный номер), используемый для ее указания как объекта оперирования при описании обработки информации.

Преобразование строки заключается в замене некоторых структурных единиц на другие строки. Таким образом, преобразование представляет собой простую подстановку вместо некоторых подстрок обрабатываемой строки других строк, и в этом смысле оно ничем не отличается от преобразований-подстановок в СНОБОЛе и РЕФАЛе. В рамках предлагаемого подхода преобразование строки распадается на следующие этапы: структурирование (разбиение строки на структурные единицы), локализацию (выделение структурных единиц, подлежащих изменению) и подстановку (изменение локализованных подстрок и образование новой строки).

Этапы преобразования относительно не зависят друг от друга, т.е. их описание связывается с разными средствами описания обработки. Так, структурирование можно считать априорным процессом, порождающим всевозможные структурные номера, локализацию - манипулированием со структурными номерами как с числовыми кортежами, а подстановку - конструированием новой строки по совокупности старых структурных единиц, номеров тех структурных единиц, которые изменяются, и заменяющих строк. Хотя из соображений эффективности такое выделение этапов преобразования не следует связывать с реализацией обработки строк, оно иногда оказывается удобным на уровне описания алгоритма.

Существенным моментом использования предлагаемого подхода является выбор функций, с помощью которых обрабатываемые строки разбиваются на структурные единицы, т.е. выделяются объекты оперирования, участвующие в обработке. От того, насколько удобным окажется структурирование для конкретных приложений, зависит адекватность использования подхода. Выбором подходящего набора функций структурирования можно добиться не только удобства и адекватности описания алгоритмов, но и наиболее эффективной реализации специализированной обработки текстов. Будет ли такая реализация адаптацией некоторой универсальной системы программирования к конкретным условиям или же специальным прикладным проектом, следует решать исходя из вида требуемых функций структурирования, частоты использования системы и т.д. В любом случае выделение подходящего набора функций структурирования позволяет решать задачу реализации специализированной текстовой обработки с большей определенностью знаний об объектах обработки.

Ниже описывается метод обработки текстовой информации, задаваемой с помощью априорного структурирования, исследуются мощ-

ность и полнота локализации, связанной со структурой. Насколько автору работы известно, предлагаемый подход и его математическое обоснование в литературе не рассматривались. В рамках настоящей статьи метод обработки текстов, связанный со структурированием, излагается в общем виде. Задача выделения классов функций структурирования для конкретных приложений и изучения мощности локализаций, использующих эти классы, может стать предметом дальнейшего исследования.

Пусть V - конечный алфавит, V^* - множество строк над V, которые мы будем называть **текстовыми значениями**, т.е. значениями, которые могут принимать тексты (содержательно под текстом можно понимать вместилище произвольных текстовых значений, с которым связан набор функций структурирования).

Атомарной единицей, которая может быть выделена в текстовом значении, является интервал.

ОПРЕДЕЛЕНИЕ 1. Пусть $\alpha = a_1 \ldots a_m$ - строка некоторого текста. Тогда тройка $\langle \beta, K1, K2 \rangle$ называется **интервалом**, если

1) $0 \le K1 < K2 \le m+1$,

2) $\beta = a_{k1+1} \cdots a_{k2-1}$;

натуральные числа K1 и K2 называются **левой и правой координатами**, а строка β - **строковой компонентой** интервала.

Если $\pi = \langle \beta, K1, K2 \rangle$ - интервал некоторой строки, то через $s(\pi), l(\pi), r(\pi)$ обозначаются его строковая компонента и координаты $\pi = \langle s(\pi), l(\pi), r(\pi) \rangle$.

Для любой строки $a_1 \ldots a_m$ существует $(m+1)(m+2)/2$ всевозможных интервалов ($m+1$ интервал вида $\langle e, K1, K1+1 \rangle$, где e - пустая строка, $0 \le K1 \le m$; m интервалов вида $\langle a_{k1+1}, K1, K1+2 \rangle$, $0 \le K1 \le m-1$; $m-1$ - вида $\langle a_{k1+1}, a_{k1+2}, K1, K1+3 \rangle$, $0 \le K1 \le m-2$ и т.д.).

ОПРЕДЕЛЕНИЕ 2. Два интервала π_1 и π_2 одной строки называются **непересекающимися**, если не выполняются условия

$$l(\pi_1) < l(\pi_2) < r(\pi_1)-1 \vee l(\pi_1) + 1 < r(\pi_2) < r(\pi_1)$$

или

$$l(\pi_2) < l(\pi_1) < r(\pi_2)-1 \vee l(\pi_2) + 1 < r(\pi_1) < r(\pi_2).$$

ОПРЕДЕЛЕНИЕ 3. Два интервала одной строки называются с м е ж н ы м и, если левая координата одного из них в точности на единицу больше правой координаты второго. Множество интервалов $\{\pi_1,\dots,\pi_n\}$ произвольной строки называется р а з б и е н и е м α, если

1) для любого $i=1,\dots,n-1$ интервалы π_i, π_{i+1} смежные;

2) $\alpha = s(\pi_1)\dots s(\pi_n)$.

Имеет место следующая

ТЕОРЕМА 1. Д л я л ю б о й с т р о к и д л и н ы $m>0$ с у щ е с т в у е т $4\cdot3^{m-1}$ р а з б и е н и й.

Для формулировки следующих определений удобно ввести частичную операцию $\ominus$ - у д а л е н и е п р е ф и к с а строки:

$$\alpha \ominus \beta = \begin{cases} \gamma_2, \text{ если } \alpha = \gamma_1\beta\gamma_2 \text{ и } \gamma_1\beta = \delta_1\beta\delta_2 \supset \delta_1 = \gamma_1 \ \& \ \delta_2 = e; \\ \qquad \gamma_1,\ \gamma_2,\ \delta_1,\ \delta_2 \in V^*; \\ \text{не определено, если } \alpha \neq \gamma_1\beta\gamma_2 \text{ для любых } \gamma_1,\gamma_2 \in V^*. \end{cases}$$

Предполагается, что $\alpha\ominus\beta'\ominus\beta'' = (\alpha\ominus\beta')\ominus\beta''$. Кроме того, полезно ввести следующие обозначения. Пусть $f: V^* \to V^*$ - частичная строковая функция. Тогда через $f[0](\alpha)$ обозначается

α_1, если $\alpha = \alpha_1 f(\alpha)\alpha_2$ и $\alpha \ominus f(\alpha) = \alpha_2$;

α, если $f(\alpha)$ определено и $\forall \alpha_1,\alpha_2 \in V^*\ \alpha \neq \alpha_1 f(\alpha)\alpha_2$.

Считается, что $f[0](\alpha)$ не определено, если $f(\alpha)$ не определено.

Если выражение $\alpha\ominus f[0](\alpha)\ominus\dots\ominus f[i-1](\alpha)$ определяет строку, к которой применима функция f, то через $f[i](\alpha)$ обозначается результат вычисления следующего рекуррентного соотношения:

$$f[i](\alpha) = f(\alpha \ominus f[0](\alpha)\ominus\dots\ominus f[i-1](\alpha))$$

при $i \geq 1$ (в противном случае $f[i](\alpha)$ не определено).

ОПРЕДЕЛЕНИЕ 4. Пусть $f: V^* \to V^*$ - частичная строковая функция, $\alpha \in V^*$ - произвольная строка, p - неотрицательное число. Тогда p-к р а т н ы м п р и м е н е н и е м ф у н к ц и и f к с т р о к е α н а з о в е м $f[p](\alpha)$, если для любого i, $0 \leq i \leq p$, $f[i](\alpha)$ определено. В противном случае будем считать, что функция f к строке α p-кратно неприменима.

Заметим, что для произвольной функции f нельзя гарантировать $f[1](\alpha) = f(\alpha)$ для любой строки $\alpha \in V^*$. Так, если для $V=\{a,b,c\}$

функция φ определяется следующим образом:

$$\varphi(\alpha) = \begin{cases} a_3, \text{ если } \alpha = a_1a_2a_3\gamma;\ a_1, a_2, a_3 \in V,\ \gamma \in V^*; \\ \text{не определено, если длина } \alpha \text{ меньше } 3, \end{cases}$$

то на строке $\gamma = x_1x_2x_3x_4x_5 \in V^*$, $x_1 \neq x_3$, $x_2 \neq x_3$, $x_5 \neq x_3$, $\varphi(\gamma) = x_3$, $\varphi[0](\alpha) = x_1x_2$, $\varphi[1](\gamma) = x_5$. Равенство $f[1](\alpha) = f(\alpha)$ выполняется при условии: если $\alpha \in V^*$, $f(\alpha)$ определено и $\alpha = \alpha_1 f(\alpha)\alpha_2$, то $f(\alpha) = f(\alpha')$ для любой строки α' вида $\beta f(\alpha)\alpha_2$, где $\beta \in V^*$ удовлетворяет соотношению $\exists\, \delta \in V^*$: $\delta\beta = \alpha_1$. Функции, для которых это условие выполняется, назовем **слабо зависимыми от префикса**. При определении структурирования текстов используются только такие функции, поэтому далее предполагается, что все рассматриваемые строковые функции слабо зависят от префикса.

ОПРЕДЕЛЕНИЕ 5. Назовем **способом разбиения** набор строковых функций $G = \{f_1, \ldots, f_n\}$, доопределенных соотношением $f(\alpha) = \omega$, $\omega \notin V$, вне областей их определений и удовлетворяющих условиям:

1) $f(\alpha) \neq \omega \supset \alpha = \alpha_1 f(\alpha)\alpha_2$ при некоторых α_1 и α_2 из V^*;

2) $\forall \alpha \in V^* \exists t \geq 0\ \forall p \in \{0, \ldots, t\}$:
$f[p](\alpha) \neq \omega \,\&\, f[t+1](\alpha) = \omega$;

3) если $f[1](\alpha) \neq \omega, \ldots, f[t](\alpha) \neq \omega$, то $\alpha = \alpha_0 f[1](\alpha)\alpha_1 \ldots \ldots \alpha_{t-1} f[t](\alpha)\alpha_t$ при некоторых $\alpha_0, \alpha_1, \ldots, \alpha_{t-1}, \alpha_t$ из V^*.

Понятно, что любая функция из способа разбиения **индуцирует** разбиение произвольной строки на интервалы со строковыми компонентами $\alpha_0, f[1](\alpha), \alpha_1, \ldots, \alpha_{t-1}, f[t](\alpha), \alpha_t$, выделяемыми условием 3 определения 5. Такое разбиение мы будем называть **каноническим**.

ОПРЕДЕЛЕНИЕ 6. Назовем **структурным разбиением** произвольной строки α, соответствующим функции f из некоторого способа структурирования, каноническое разбиение, из которого удалены все интервалы вида $\langle \alpha_i, K1_i, K2_i \rangle$, $0 < i < t$, если $\alpha_i = 1$.

ОПРЕДЕЛЕНИЕ 7. Если функция f из некоторого способа разбиения G такова, что структурное разбиение любой строки α, соответствующее f, состоит из интервалов со строковыми компонентами $\alpha_0, f[1](\alpha), f[2](\alpha), \ldots, f[t](\alpha), \alpha_t$ (т.е. $\alpha_1 = \alpha_2 = \ldots = \alpha_{t-1} = e$ в условии 3 определения 5), то f называется **правильно разбивающей функцией** или **функцией струк-**

турирования. Если любая функция способа разбиения является функцией структурирования, то такой способ называется способом структурирования.

Для произвольной строки α и функции структурирования f число интервалов h структурного разбиения удовлетворяет условию $t \le h \le t+1$, где $t = \max p(f[p](\alpha) \neq \omega)$. Легко показать, что для любой функции f из произвольного способа разбиения G выполняется

$$(\alpha \neq e \;\&\; f[i](\alpha) \neq \omega) \supset f[i](\alpha) \neq e.$$

Непосредственно из определения способа разбиения следует, что для любой функции f из какого-либо способа разбиения справедливо $f(e) = \omega$. В частности, функция с множеством неподвижных точек, совпадающим с V^*, не может принадлежать никакому способу разбиения. Напротив, функция, определяемая соотношением $\forall \alpha \in V^*: (\alpha \neq e \supset f(\alpha) = \alpha) \,\&\, (\alpha = e \supset f(\alpha) = \omega)$, удовлетворяет условиям I–3 определения 5 и, более того, является функцией структурирования. Функциями структурирования являются также все функции, для которых справедливо $\forall \alpha \in V^*: (f(\alpha) \neq \omega) \supset \exists \beta \in V^*: (\alpha = f(\alpha)\,\beta)$. Этот класс функций очень важен с практической точки зрения.

Следующий результат обосновывает достаточность введенного определения функций структурирования для практических целей.

ТЕОРЕМА 2. Для любой функции t какого-либо способа структурирования (т.е. удовлетворяющей условиям I–3 определения 5) существует правильно разбивающая функция $\tilde{f}$, любое структурное разбиение которой совпадает с каноническим для f.

ДОКАЗАТЕЛЬСТВО. Определим $\tilde{f}$ следующим образом:

$$\tilde{f}(\alpha) = \begin{cases} f(\alpha), \text{если } \alpha = f(\alpha)\beta_2,\ \beta_2 \in V^*; \\ \beta_1, \text{ если } \alpha = \beta_1 f(\alpha)\beta_2 \,\&\, \beta_1 \neq e \,\&\, (\beta_1 f(\alpha) = \\ \qquad = \beta_1' f(\alpha)\ \beta_2' \supset \beta_1' = \beta_1 \,\&\, \beta_2' = e),\ \beta_1, \beta_2 \in V^*; \\ \text{не определено, если } f(\alpha) \text{ не определено.} \end{cases}$$

Учитывая, что все рассматриваемые функции слабо зависят от префикса, имеем: $\forall \alpha \in V^* \forall i: f[i](\alpha) \neq \omega \;\exists j: \tilde{f}[j](\alpha) = f[i](\alpha)$. Кро-

ме того, для любой строки, такой, что $f(\alpha)$ определено $\tilde{f}(\alpha)$ удовлетворяет условию $\alpha = \tilde{f}(\alpha)\gamma$, $\gamma \in V^*$. Последнее утверждение завершает доказательство.

По произвольному разбиению с непустыми строковыми компонентами интервалов $\Pi_\alpha = \{\pi_1, \ldots, \pi_n\}$ некоторой строки α можно построить функцию структурирования f^{Π_α}, структурирование которой для строки α будет совпадать с Π_α. Для строки $\beta \in V^*$, $\beta \neq \alpha$, f^{Π_α} определяется следующим образом:

$$f^{\Pi_\alpha}(\beta) = \begin{cases} s(\pi_1), \text{ если } \beta = \alpha_1\beta_2,\ \alpha = \alpha_1\alpha_2,\ \beta_2 \neq \alpha_2 \text{ и} \\ \qquad \alpha_1 = s(\pi_1)\gamma,\ \alpha_1, \alpha_2,\ \beta_2,\ \gamma \in V^*; \\ s(\pi_2), \text{ если } \beta = \alpha_1\beta_2,\ \alpha \ominus s(\pi_1) = \alpha_1\alpha_2,\ \beta_2 \neq \alpha_2 \text{ и} \\ \qquad \alpha_1 = s(\pi_2)\gamma,\ \alpha_1, \alpha_2,\ \beta_2,\ \gamma \in V^*; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ s(\pi_n), \text{если } \beta = \alpha_1\beta_2,\ \alpha \ominus s(\pi_1) \ominus \ldots \ominus s(\pi_{n-1}) = s(\pi_n) = \alpha_1, \\ \qquad \beta_2 \neq e,\ \alpha_1,\ \beta_2 \in V^*; \\ \text{не определено - в остальных случаях.} \end{cases}$$

Функцию f^{Π_α} будем называть индуцированной разбиением Π_α строки α.

Понятно, что f^{Π_α} является функцией структурирования; далее, структурирование α совпадает с разбиением Π_α, для $\beta = \beta_1 s(\pi_i) \ldots s(\pi_j)\beta_2$, $1 \le i \le j \le n$, структурирование приводит к получению разбиения $\langle \beta_1, K1_1, K2_1\rangle, \langle s(\pi_i), K1_2, K2_2\rangle, \ldots, \langle s(\pi_j), K1_{j-i+2}, K2_{j-i+2}\rangle, \langle \beta_2, K1_{j-i+3}, K2_{j-i+3}\rangle$ при подходящих значениях $K1_h$, $K2_h$, $1 \le h \le j-i+3$. Далее, для любого набора непересекающихся интервалов Π_α произвольной строки α легко построить разбиение α с минимальным количеством интервалов, содержащее Π_α, и, следовательно, справедлива

ТЕОРЕМА 3. Для любого значения α текста t произвольный набор Π непересекающихся интервалов с непустыми строковыми компонентами индуцирует функцию структурирования f_t^{Π} такую, что структурное разбиение α, по-

лучаемое с помощью этой функции, есть стандартное для α и Π.

Для любой строки α (значения некоторого текста t) и допустимого значения $p \geq 0$ p-кратное применение функции структурирования выделяет в α некоторый интервал. Процесс выделения интервала строки будем называть локализацией. В принципе, "локализующая способность" функций структурирования достаточно велика: выбирая подходящим образом их определение, можно добиться локализации любого непустого интервала в конкретной строке. Однако не следует забывать, что вся сложность задания текстовой обработки, нисколько не уменьшаясь, перекладывается в таком случае на определение функций локализации. Для увеличения локализующей способности функций структурирования необходимо использование некоторого механизма взаимодействия вычислений этих функций на обрабатываемых строках. Следующая наша цель состоит в том, чтобы определить такой механизм.

ОПРЕДЕЛЕНИЕ 8. Пусть $G = \{f_1, \ldots, f_s\}$ - способ структурирования, $\alpha = a_1, \ldots, a_m$ - произвольная строка из V^*. Структурным номером называется кортеж из h целых неотрицательных чисел, $0 \leq h \leq n$. Единственной структурной единицей уровня 0 является интервал $\langle \alpha, 0, m+1 \rangle$, выделяемый пустым структурным номером ($h = 0$). Пусть $\langle \sigma_{h-1}, K1_{h-1}, K2_{h-1} \rangle$ есть структурная единица уровня $h-1$, $0 < h \leq n$, выделяемая в α с помощью структурного номера $(p_1, \ldots, p_{h-1})$. Это означает, что α представима в виде $\alpha'_{h-1}\sigma_{h-1}\alpha''_{h-1}$. Пусть далее $\alpha_h = \sigma_{h-1}\alpha''_{h-1}$. Тогда структурная единица уровня h, выделяемая структурным номером $(p_1, \ldots, p_{h-1}, p_h)$, есть интервал $\langle \sigma_h, K1_h, K2_h \rangle$, в котором $\sigma_h = f_h[p_h](\alpha_h)$, а $K1_h, K2_h$ - координаты в α подстроки σ_h, если $f_h[1](\alpha_h) \neq \omega, \ldots, f_h[p_h](\alpha_h) \neq \omega$. В противном случае структурный номер $(p_1, \ldots, p_h)$ не выделяет структурной единицы в текстовом значении α.

Через $(p_1 \bullet \ldots \bullet p_h)(\alpha)$ будем обозначать интервал, выделяемый структурным номером $(p_1, \ldots, p_h)$ по некоторому способу структурирования в некоторой строке α. Если с произвольным текстом связано несколько способов структурирования, то мы будем обозначать структурный номер и соответствующую структурную единицу, относящиеся к некоторому конкретному способу структурирования G, соответственно через $(p_1, \ldots, p_h)_G$ и $(p_1 \bullet \ldots \bullet p_h)_G(\alpha)$. Чтобы подчеркнуть зависимость структурирующих функций способа структурирования по отношению к последнему будем иногда употреблять термин иерар-

хическая группа функций структурирования.

Иерархическим структурированием определяется специальный класс разбиений текстовых значений. Если $G = \{f_1,\ldots,f_n\}$ - способ структурирования текста t, а α - его произвольное значение, такое, что $(p_1,\ldots, p_{h-1})$ выделяет в α структурную единицу, $1\le h \le n$, то функция структурирования f_h задает правильное разбиение (структурирование) строки $\gamma = \gamma_1\gamma_2$, где $\gamma_1 = s((p_1 \circ \ldots \circ p_{h-1})_G(\alpha))$ и $\gamma_2 = \alpha \ominus s((p_1 \circ \ldots \circ p_{h-1})_G(\alpha))$. Если к тому же $p_1=\ldots=p_{h-1} = 0$, то f_h структурирует все значение α точно так, как и без использования иерархического структурирования. Последнее справедливо для любого $h=1,\ldots,n$ (случай $h=1$ тривиален). Поскольку для любого h, $0\le h\le n$, структурный номер $(\underbrace{0,\ldots,0}_{h})$ всегда выделяет некоторую структурную единицу произвольного текстового значения, нами доказана

ЛЕММА I. Мощность множества интервалов, выделяемых неиерархическим структурированием текста с помощью набора функций $f_1,\ldots,f_n$, не превосходит мощности множества интервалов, выделяемых иерархическим структурированием, использующим эти же функции.

Следующий пример показывает, что лемма может быть усилена. Пусть φ: $\{a,b,c\}^* \to \{a,b,c\}^* \cup \{\omega\}$ - функция структурирования некоторого текста t со значениями из $\{a,b,c\}^*$, которая определяется таким образом:

$$\varphi(\alpha) = \begin{cases} \beta a, & \text{если } \alpha=\beta a\gamma \text{ и } \beta \text{ не содержит вхождений символа a;} \\ \alpha, & \text{если для любых } \beta \text{ и } \gamma \ \alpha\neq\beta a\gamma \text{ и } \alpha\neq e; \\ \omega, & \text{если } \alpha = e. \end{cases}$$

Аналогично определим функцию ϕ:

$$\phi(\alpha) = \begin{cases} \beta b, & \text{если } \alpha=\beta b\gamma \text{ и } \beta \text{ не содержит вхождений символа b;} \\ \alpha, & \text{если для любых } \beta \text{ и } \gamma \ \alpha\neq\beta b\gamma \text{ и } \alpha\neq e; \\ \omega, & \text{если } \alpha = e. \end{cases}$$

Очевидно, что φ и ϕ - функции структурирования, однако интервал $(2.I)_{\{\varphi,\phi\}}$ (ccabcccab) = ⟨b,3,5⟩ не может быть получен с помощью неиерархического структурирования. Таким образом, справедлива

ТЕОРЕМА 4. Множество интервалов, выделяемых неиерархическим структурированием, включается в множество всех иерархических интервалов, выделяемых теми же самыми функциями структурирования.

Часто под иерархией структуры текста понимают вложенность структурных единиц. Наше определение не соответствует такому пониманию, однако вложенность, задаваемая разными функциями структурирования, легко моделируется с помощью иерархий и операций над интервалами. К их определению мы сейчас приступим. Вложенность же структурных единиц, выделяемых с помощью одной и той же функции структурирования, может быть задана с помощью языковых средств вычислительного характера и структур управления алгоритмического языка, которые в настоящей работе не рассматриваются.

Наше структурирование имеет большую аналогию с измерениями и системами мер, например, для нахождения длин отрезков. С любой степенью точности длина отрезка может быть представлена в виде n чисел: p_1 километров, p_2 метров, p_3 дециметров и т.д. - (p_1,... , p_n). Если измерение не включает, например, "километров", то первая компонента такого набора может считаться нулевой: (0,p_2,... , p_n). Весь отрезок можно считать разбитым на интервалы,соответствующие любым допустимым значениям p_1,...,p_n точно так же, как текстовое значение разбивается на структурные единицы. Однако эта аналогия неполная, поскольку в случае структурирования,образно говоря, разные "километры" могут содержать разное количество "метров"; структурная единица не точка (нечто не имеющее внутренних размеров), а интервал, с которым можно оперировать как с обыкновенной строкой.

Для введения гибких средств задания локализуемых участков недостаточно простого структурирования текстовых значений. Одной из возможностей увеличения мощности этих средств является введение операций, объединяющих локализованные структурированием структурные единицы, которые мы будем называть операциями текстовых локализаций или объединениями интервалов.

ОПРЕДЕЛЕНИЕ 9. Пусть π_1 и π_2- некоторые интервалы текстового значения $\alpha = a_1 \ldots a_m$. Тогда, если вводимые ниже интервалы определены корректно,

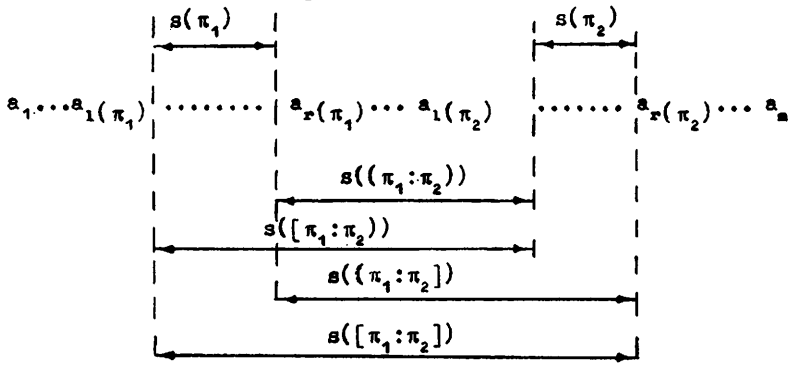
$[\pi_1 : \pi_2] = \langle \gamma^1, l(\pi_1), r(\pi_2) \rangle$ - замкнутое объединение π_1 и π_2, $(\gamma^1 = a_{l(\pi_1)+1} \cdots a_{r(\pi_2)-1})$;

$(\pi_1 : \pi_2) = \langle \gamma^2, r(\pi_1)-1, l(\pi_2)+1 \rangle$ - открытое объединение π_1 и π_2, $(\gamma^2 = a_{r(\pi_1)} \cdots a_{l(\pi_2)})$;

$(\pi_1 : \pi_2] = \langle \gamma^3, r(\pi_1)-1, r(\pi_2) \rangle$ - открытое слева объединение π_1 и π_2, $(\gamma^3 = a_{r(\pi_1)} \cdots a_{r(\pi_2)-1})$;

$[\pi_1 : \pi_2) = \langle \gamma^4, l(\pi_1), l(\pi_2)-1 \rangle$ - открытое справа объединение π_1 и π_2, $(\gamma^4 = a_{l(\pi_1)+1} \cdots a_{l(\pi_2)})$.

Для пояснения определения 9 на рисунке приведен пример конкретного использования операций объединения.



Заметим, что операции объединения определены не для любых пар интервалов, поэтому, строго говоря, они являются частичными операциями. Однако мы не будем впредь употреблять слово "частичные" по отношению к этим операциям, понимания это неявно. Условия корректности операций текстовых локализаций приведены в таблице.

Таблица

Вид операции	Условие корректности
$[\pi_1 : \pi_2]$	$l(\pi_1) < r(\pi_2)$
$(\pi_1 : \pi_2)$	$r(\pi_1)-1 < l(\pi_2)+1$
$[\pi_1 : \pi_2)$	$l(\pi_1) < l(\pi_2)+1$
$(\pi_1 : \pi_2]$	$r(\pi_1)-1 < r(\pi_2)$

Набор операций объединения избыточен для задания всех видов объединений, поскольку достаточно ввести, например, операции "[:]" и "(:)". Так, если для каких-либо интервалов π' и π'' определены $[\pi' : \pi'')$ или

($\pi':\pi"$], то их можно выразить с помощью суперпозиций "[:]" и "(:)":

$$[\ \pi':\pi"\ ) = [\ \pi':(\pi':\pi")],$$

$$(\ \pi':\pi"\ ] = [\ \pi':\pi"):\pi"].$$

Если с некоторым текстом t связано u иерархических групп функций структурирования $G_1,\dots,G_u$, то может оказаться определенным результат операций объединения интервалов, выделяемых с помощью иерархического структурирования $(p_i^1 \bullet \dots \bullet p_i^q)_{G_i}(\alpha)$ и $(p_j^1 \bullet \dots \bullet p_j^{q'})_{G_j}(\alpha)$, $i \le i$, $j \le u$, $q \le h_i$, $q' \le h_j$, где h_i, h_j - число функций структурирования в группах G_i и G_j, соответственно. Определим $\mathcal{M}(G_1,\dots,G_u;\alpha)$ как множество всех интервалов α, получаемых с помощью суперпозиций операций объединения интервалов, выделяемых всевозможными структурными номерами $(p_i^1,\dots,p_i^q)_{G_i}$, $1 \le i \le u$, $0 \le q \le h_i$, где h_i - число функций структурирования группы G_i.

Следующие результаты касаются сравнения мощностей локализаций с помощью иерархического структурирования и неиерархического, если они снабжены операциями объединения.

ТЕОРЕМА 6. Пусть t-текст, с которым связано u групп функций структурирования $G_1,\dots,G_u$, $G_i = \{f_i^1,\dots,f_i^{h_i}\}$, $1 \le i \le u$. Тогда для любого значения α текста t справедливо

$$\mathcal{M}(G_1,\dots,G_u;\alpha) \supseteq \mathcal{M}(\{f_1^1\},\dots,\{f_1^{h_1}\},\dots,\{f_u^1\},\dots,\{f_u^{h_u}\};\alpha).$$

ДОКАЗАТЕЛЬСТВО непосредственно следует из леммы I.

ТЕОРЕМА 7. Для произвольного текста t существует набор функций структурирования $f^1,\dots,f^h$ такой, что

$$\bigcup_{\alpha \in V^*} \mathcal{M}(\{f^1\},\dots,\{f^h\};\alpha) \subsetneq \bigcup_{\alpha \in V^*} \mathcal{M}(\{f^1,\dots,f^h\};\alpha).$$

ДОКАЗАТЕЛЬСТВО теоремы сводится к построению такого набора функций, который на некотором текстовом значении α может породить $\pi \in \mathcal{M}(\{f^1,\dots,f^h\};\alpha)$ и $\pi \notin \mathcal{M}(\{f^1\},\dots,\{f^h\};\alpha)$.

Пусть с некоторым текстом t связано u способов структурирования $G_1,\ldots,G_u$. Обозначим через $\mathcal{M}_0(G_1,\ldots,G_u;\alpha)$ множество всех структурных единиц текстового значения t, через $\mathcal{M}_{i+1}(G_1,\ldots,G_u;\alpha)$ – множество всех интервалов, получаемых с помощью однократного вычисления операций текстовых локализаций на $\mathcal{M}_i(G_1,\ldots,G_u;\alpha)$, $i = 0,1\ldots$. Очевидно, $\mathcal{M}(G_1,\ldots,G_u;\alpha) = \bigcup_{i=0}^{\infty} \mathcal{M}_i(G_1,\ldots,G_u;\alpha)$. Следующий результат касается соотношения мощностей всех этих множеств и показывает, что использование суперпозиций текстовых локализаций избыточно для задания любого интервала из $\mathcal{M}(G_1,\ldots,G_u;\alpha)$.

ТЕОРЕМА 8. Пусть $G_1,\ldots,G_u$ – способы структурирования некоторого текста t. Тогда для любого текстового значения справедливо

$$\mathcal{M}_0(G_1,\ldots,G_u;\alpha) \subset \mathcal{M}_1(G_1,\ldots,G_u;\alpha),$$

$$\mathcal{M}_1(G_1,\ldots,G_u;\alpha) = \mathcal{M}_2(G_1,\ldots,G_u;\alpha) = \ldots = \mathcal{M}(G_1,\ldots,G_u;\alpha).$$

Кроме того, для любого нетривиального набора $G_1,\ldots,G_u$, т.е. выделяющего хотя бы две структурные единицы с непустыми и несовпадающими строковыми компонентами, существует такая строка α, что

$$\mathcal{M}_0(G_1,\ldots,G_u;\alpha) \neq \mathcal{M}_1(G_1,\ldots,G_u;\alpha).$$

ДОКАЗАТЕЛЬСТВО. Для краткости вместо $\mathcal{M}_i(G_1,\ldots,G_u;\alpha)$ и $\mathcal{M}(G_1,\ldots,G_u;\alpha)$ будем писать $\mathcal{M}_i(\alpha)$ и $\mathcal{M}(\alpha)$ или даже $\mathcal{M}_i$ и $\mathcal{M}$. Рассмотрим каждое из утверждений теоремы отдельно.

I. $\mathcal{M}_0 \subset \mathcal{M}$. Тривиально следует из определения замкнутого объединения интервалов.

II. $\mathcal{M}_1 = \mathcal{M}_2 = \ldots = \mathcal{M}$. Очевидна следующая цепочка включений: $\mathcal{M}_1 \subseteq \mathcal{M}_2 \subseteq \ldots \subseteq \mathcal{M}$. Обратная цепочка доказывается по индукции.

а) $\mathcal{M}_1 \subseteq \mathcal{M}_2$. Пусть π – произвольный интервал из $\mathcal{M}_2$. Это означает, что $\pi = [\pi_1:\pi_2]$, либо $\pi = (\pi_1:\pi_2)$, либо $\pi = [\pi_1:\pi_2)$, либо $\pi = (\pi_1:\pi_2]$, где π_1 или π_2 представим как результат одной из операций текстовой локализации. Рассмотрение каждого из возможных случаев доказывает требуемое включение. Например, для $\pi = [(\pi_{11}:\pi_{12}]:\pi_2)$ имеем $\pi = (\pi_{11}:\pi_2) \in \mathcal{M}_1$.

б) Пусть $\mathcal{M}_1 \supseteq \ldots \supseteq \mathcal{M}_{k-1}$. Докажем, что $\mathcal{M}_{k-1} \supseteq \mathcal{M}_k$, $k > 2$. Для произвольного $\pi \in \mathcal{M}_k$ имеем $\pi = \{\pi_1 : \pi_2\}$, где "{" и "}" обозначают одну из возможных скобок, идентифицирующих операцию. Тогда существует i такое, что $\pi_1 \in \mathcal{M}_i$, а $\pi_2 \in \mathcal{M}_{k-i-1}$, $0 \le i \le k$. Так как $\max(i, k-i-1) < k$, по предположению индукции, $\pi_1 \in \mathcal{M}_1$ и $\pi_2 \in \mathcal{M}_1$, и, следовательно, $\pi \in \mathcal{M}_2$, что завершает доказательство утверждения П.

Ш. $\exists \alpha : \mathcal{M}_0(\alpha) = \mathcal{M}(\alpha)$. Пусть в некоторой строке α выделяются структурные единицы π_1 и π_2, такие, что $s(\pi_1) \neq e$, $s(\pi_2) \neq e$ и $s(\pi_1) \neq s(\pi_2)$. Из этого следует, что $l(\pi_1) \neq l(\pi_2)$ или $r(\pi_1) \neq r(\pi_2)$. Рассмотрим первый случай; пусть для определенности $l(\pi_1) < l(\pi_2)$. Интервал $\pi_3 = [\pi_1 : \pi_2)$ корректен, и $l(\pi_3) = l(\pi_1)$, $r(\pi_3) = l(\pi_2) + 1$. Пусть, далее, $\pi_3 \in \mathcal{M}_0(\alpha)$. Тогда интервал $\pi = (\pi_3 : \pi_2)$ не может содержаться в $\mathcal{M}_0(\alpha)$, так как $s(\pi) = e$, $l(\pi) = r(\pi_3) - 1 = l(\pi_2) > l(\pi_1)$, поэтому $l(\pi) > 0$ и, следовательно, $l(\pi) = m$, $r(\pi) = m+1$, но $r(\pi) = l(\pi_2) + 1$, что невозможно из-за $s(\pi_2) \neq e$. Для случая $r(\pi_1) \neq r(\pi_2)$ доказательство аналогично.

Описанный выше механизм структурирования текстов был положен в основу средств локализации участков текстов в процессоре Глобальной Текстовой Обработки [4], реализованном на ЭВМ БЭСМ-6. Использование процессора при решении различных задач обработки текстов показало удобство средств конструирования структурирования строковой информации и потенциальную возможность их эффективной реализации.

Литература

I. GRISWOLD R.E. at al. The SNOBOL 4 Programming Language. - Prentice. Hall, Englewood Cliffs, N.Y., 2 nd.ed., 1971.

2. ТУРЧИН В.Ф. Базисный РЕФАЛ. Описание языка и основные приемы программирования. (Методические рекомендации).-М.: ЦНИИАСС, 1974, с.96.

3. МАРКОВ А.А. Теория алгорифмов. -Л.-М.: Изд-во Акад. наук, СССР, 1954, 375 с. (Тр. матем. ин-та АН СССР, т.42).

4. ГЕЙДАН Л.И., СКОПИН И.Н. Процессор ГТО как средство автоматизации обработки текстов. -В кн.: Математическое обеспечение моделирования сложных систем. Ч.I, Киев, 1977, с.180-182.

Поступила в ред.-изд.отд.
13 марта 1980 года