АКАДЕМИЯ НАУК СССР
СИБИРСКОЕ ОТДЕЛЕНИЕ
ИНСТИТУТ МАТЕМАТИКИ

СТРУКТУРНЫЙ АНАЛИЗ СИМВОЛЬНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

(Вычислительные системы, 101)

Сборник научных трудов

Научные редакторы:

доктор технических наук Ю. Г. КОСАРЕВ
кандидат технических наук В. Д. ГУСЕВ

НОВОСИБИРСК 1984

СТРУКТУРНЫЙ АНАЛИЗ СИМВОЛЬНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ
(Вычислительные системы)
1984 год Выпуск 101

УДК 519.766

ПАКЕТ ПРИКЛАДНЫХ ПРОГРАММ ДЛЯ АНАЛИЗА ПРОИЗВОЛЬНЫХ СИМВОЛЬНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ ЗНАЧИТЕЛЬНОЙ ДЛИНЫ (СИМВОЛ)

В.Д.Гусев, Ю.Г.Косарев, М.К.Тимофеева, Т.Н.Титкова,
Н.А.Чужанова, Г.С.Высоцкая

I. Введение

Пакет прикладных программ для анализа произвольных символьных последовательностей (текстов) значительной длины (ППП СИМВОЛ) ориентирован на выявление структурных и частотных свойств текстов. Основные методы и алгоритмы, применяемые в пакете, являются оригинальными разработками.

С помощью пакета пользователи могут проводить достаточно широкий круг исследований символьных последовательностей. В частности, пакет может использоваться как для первоначального знакомства с неизвестными текстами, предоставляя на суд исследователя их основные структурные и частотные закономерности, так и для детального исследования заинтересовавших его характеристик текста и отдельных его элементов.

Объектом исследования могут быть тексты на естественных языках, первичные структуры нуклеотидных молекул, последовательности типов пород, вскрываемых при бурении скважин, тексты программ и т.п.

ППП СИМВОЛ ориентирован на массового пользователя, не являющегося профессиональным программистом. В нем предусмотрены специальные языковые и программные средства для упрощения как обращения к пакету, так и организации ввода/вывода данных. Процесс программирования сведен к сборке уже готовых модулей из библиотеки подпрограмм пакета по описаниям исходных и результирующих данных.

ППП СИМВОЛ работает под управлением операционной системы ОС ЕС в режимах РСР, МФТ или МУТ на ЭВМ ЕС-1022 и старше с объемом оперативной памяти 512 К.

ППП СИМВОЛ состоит из транслятора с языка пользователя пакетом, библиотек обрабатывающих модулей и сервисных программ, управляющей программы пакета.

2. Методы исследования текстов

1. Термины и обозначения.

А л ф а в и т - конечное множество символов; |A| - мощность алфавита А. Среди алфавитов особое место занимает а л ф а в и т т е р м и н а л ь н ы х с и м в о л о в (термов), в котором принято выражать результаты решения задач. Далее, как это делается обычно, в качестве такового берется множество символов, каждому из которых соответствует графический знак на стандартных устройствах печати ЭВМ (АЦПУ и пишущие машинки типа "Консул").

Среди нетерминальных символов важную роль играют так называемые с и н т е р м ы, которые представляют собой имена некоторых выделенных подмножеств терминального алфавита. Использование синтермов сокращает описание данных и запись алгоритмов там, где различные термы из некоторого их подмножества выступают в одинаковой роли. Примеры синтермов: цифра: ЦА:=0,1,2,...,9; разделитель: # := ".", ",", ";", "!", "?", ":", "␣".

Т е к с т в а л ф а в и т е А - конечная непустая последовательность символов из А; T[i] - i-й элемент текста T ($1 \le i \le N$, где N - д л и н а текста); T[i:j] - элементы текста с i-го по j-й включительно ($1 \le i < j \le N$).

Ц е п о ч к а в а л ф а в и т е А означает то же, что и "текст", но может не содержать ни одного символа (Λ - пустая цепочка). Обычно используется для обозначения части текста или последовательностей небольшой длины.

К о н к а т е н а ц и я ц е п о ч е к (текстов) $X_1 = a_1a_2...a_{N1}$ и $X_2 = b_1b_2...b_{N2}$ есть цепочка (текст) $X = a_1a_2...a_{N1}a_{N1+1}...a_N$ длиной $N = N1 + N2$, где $a_{N1+i} = b_i$, $i = 1,...,N2$. Для обозначения конкатенации будет использоваться знак ".", например, $X_1 \cdot X_2$.

l - г р а м м а - связная подпоследовательность текста из l подряд расположенных символов ($l = 1,2,..., N$). В тексте длины N содержатся M_l различных l-грамм ($1 \le M_l \le N-l+1$); $M_l = 1$ соот-

ветствует тексту, образованному повторением одного и того же символа, а $M_1 = N-1+1$ - тексту без повторяющихся 1-грамм.

$l_{max}(T)$ - наибольшее значение l, при котором в тексте T еще содержатся повторяющиеся 1-граммы.

Частотная характеристика текста T порядка 1 - совокупность элементов $\Phi_1(T) = \{\varphi_{1\,1}, \ldots, \varphi_{1M_1}\}$, где каждый элемент φ_{1i} $(1 \le i \le M_1)$ есть пара ⟨i-я 1-грамма, частота ее встречаемости в тексте⟩. Естественно, что для $l > l_{max}$ все 1-граммы будут единичными, т.е. будут встречаться по одному разу.

Полный частотный спектр текста T - совокупность частотных характеристик $\Phi(T) = \{\Phi_1(T), \ldots, \Phi_{l_{max}}(T); \Phi_{l_{max}+1}(T); \Phi_{l_{max}+2}(T)\}$. По $\Phi_{l_{max}+2}(T)$ (списку всех единичных $(l_{max}+2)$-грамм) текст T может быть восстановлен однозначно с помощью простого алгоритма нахождения пар $(l_{max}+2)$-грамм с совпадающими $(l_{max}+1)$-граммами.

На практике обычно имеют дело с неполными частотными спектрами текста T. Ограничения могут касаться как набора длин 1-грамм, так и самих 1-грамм путем задания их списка и/или указания пороговых значений частот их встречаемости F_1 $(\check{F}_1 \le F_1 \le \hat{F}_1)$.

Интегральные (вторичные) характеристики текста - совокупность числовых параметров, получаемых на основе частотных (первичных) характеристик текста. Примеры интегральных характеристик: M_1; l_{max}; F_1^{max} - максимальная частота встречаемости среди всех 1-грамм, входящих в $\Phi_1(T)$; E_1^k - количество различных 1-грамм в $\Phi_1(T)$, каждая из которых встречается в тексте ровно k раз $(k = 0,1,2,\ldots, N-1+1)$.

Совместная частотная характеристика 1-го порядка текстов T_1 и T_2 - совокупность элементов $\Phi_1(T_1,T_2) = \{\varphi_{1\,1}(T_1,T_2), \varphi_{1\,2}(T_1,T_2), \ldots, \varphi_{1M_1(T_1,T_2)}(T_1,T_2)\}$, где $M_1(T_1,T_2)$ - количество различных 1-грамм, общих для обоих текстов, $(0 \le M_1(T_1,T_2) \le \min\{M_1(T_1), M_1(T_2)\})$, а каждый элемент $\varphi_{1i}(T_1,T_2)$ $(1 \le i \le M_1(T_1,T_2))$ есть тройка ⟨i-я 1-грамма, частота ее встречаемости в T_1, частота ее встречаемости в T_2⟩.

Полный совместный частотный спектр текстов T_1 и T_2 - совокупность совмест -

ных частотных характеристик $\Phi(T_1,T_2)=\{\Phi_1(T_1,T_2),\ \Phi_2(T_1,T_2),\ldots,\ \Phi_L(T_1,T_2)\}$, где L - максимальное значение l, при котором в текстах T_1 и T_2 еще есть общие l-граммы (т.е. $M_L(T_1,T_2)\neq 0$, а $M_{L+1}(T_1,T_2)=0$).

Как и в случае одного текста, обычно имеют дело с неполными спектрами, у которых существуют ограничения на длины l-грамм и/или на список самих l-грамм и/или на диапазон частот их встречаемости.

U является **подпоследовательностью** текста T, если существует монотонно возрастающая последовательность целых $r_1,r_2,\ldots,r_{|U|}$, такая, что $U[i]=T[r_i]$ для $1\le i\le |U|$. U является **общей** подпоследовательностью текстов T_1 и T_2, если U - подпоследовательность как T_1, так и T_2. **Максимально длинная общая подпоследовательность** (МДП) есть общая подпоследовательность с наибольшим возможным числом элементов. К примеру, тексты T_1= abcbdda и T_2= badbabd имеют МДП U = abbd (выделена подчеркиванием) длиной $p(T_1,T_2)=4$.

Редакционное расстояние между текстами T_1 и T_2 - $d(T_1,T_2)$ - наименьшее число шагов, требующихся для перевода одной последовательности в другую, где под шагом понимается любая из следующих элементарных операций:

а) замена одного символа другим;

б) устранение одного символа из текста;

в) включение одного символа в текст.

Каждая из этих операций может иметь свою "стоимость". Если принять стоимость операций "б" и "в" равной I, а стоимость операции "а" - 2, то длина $p(T_1,T_2)$ МДП будет связана с введенным таким образом расстоянием $d(T_1,T_2)$ соотношением $d(T_1\ T_2)=|T_1|+|T_2|-2p(T_1,T_2)$.

Коэффициент конкордации - мера согласованности независимых упорядочений R объектов m экспертами. Может изменяться в диапазоне от 0 (упорядочения несогласованны) до I (идентичные упорядочения).

Коэффициент конкордации текстов $T_1, T_2,\ldots,T_m$ 1-го порядка - мера сходства (близости) m текстов по частотным характеристикам 1-го порядка. Каждый текст представляется частотной характеристикой 1-го порядка. Все l-граммы, вхо-

дящие в $\Phi_1(T)$, упорядочиваются по убыванию частоты встречаемости. Потенциально возможное число 1-грамм $k_1 = n^1$ (число упорядочиваемых объектов). Тем из них, которые отсутствуют в $\Phi_1(T)$,присваивается нулевая частота. Коэффициент конкордации m полученных упорядочений n^1 объектов характеризует меру близости текстов по частотным характеристикам 1-го порядка.

О р г р а ф G(V,E) - непустое множество вершин V и множество дуг E, представленных упорядоченными парами вершин (v,w). Вершина v называется н а ч а л о м, w - к о н ц о м дуги (v, w) . П у т е м из вершины v_1 в вершину v_2 называется последовательность вершин $v_1, v_2, \ldots, v_n$, соединенных дугами (v_1, v_2) , $(v_2, v_3), \ldots, (v_{n-1}, v_n)$.

Д е р е в о - орграф, у которого имеется в точности одна вершина, называемая в х о д о м (или к о р н е м).В эту вершину не входит ни одна дуга, а из нее к каждой вершине ведет единственный путь. В каждую вершину, не являющуюся корнем, входит ровно одна дуга. Вершины, у которых нет выходных дуг, называются терминальными (или листьями). Дуги деревьев условимся метить символами из алфавита A. Тогда каждому пути $(v_1, v_2, \ldots, v_n)$ на дереве будет взаимно-однозначно соответствовать цепочка символов из A $(a_1, \ldots, a_{n-1})$, а дереву t_Z будет взаимно-однозначно соответствовать множество цепочек Z.

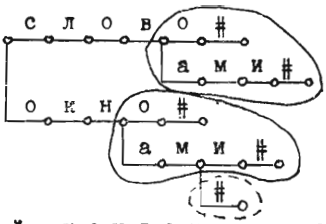
П о д д е р е в о, корнем которого является вершина v_0 дерева t, есть дерево, содержащее все пути, выходящие из вершины v_0.

Л и н е й н ы й у ч а с т о к д е р е в а - такое поддерево, из каждой нетерминальной вершины которого выходит ровно одна дуга.

У п а к о в к а ц е п о ч е к в д е р е в о - процесс построения дерева по заданному множеству цепочек.

В алфавите A выделим подмножество R символов разделителей. Если заменить каждый символ из R символом $\# \notin A$, то текст T разобьется на последовательность цепочек $w_1 \#, w_2 \#, \ldots, w_n \#$, которые назовем с е г м е н т а м и текста T. Исключив из этой последовательности повторяющиеся сегменты, получим множество сегментов W . Соответствующее множеству W дерево обозначим через t_W.

ПРИМЕР I.Множество сегментов W = {"слово#", "словами #", "окно #", "окнами #", "окнам#"} представимо в виде дерева t_W:



Левый контекст цепочки e – цепочка Δ такая, что $\Delta e\#$ есть сегмент, т.е. $\Delta e\# \in W$; K_e – множество всех левых контекстов для e.

Цепочки e' и e связаны отношением взаимозамещаемости τ_s, если у них не менее s общих левых контекстов $|K_e \cap K_{e'}| \geq s$, где s – заданная целочисленная константа. Контексты из $K_e \cap K_{e'}$ в этом случае будем называть контекстами взаимозамещаемости для e и e'.

В примере I цепочки "о" и "ами" связаны отношением взаимозамещаемости τ_2, так как у них имеются два общих левых контекста "слов" и "окн".

Пусть сегменту $a_1 \ldots a_m b_1 \ldots b_{n-1}\# \in W$ соответствует путь $u_1, \ldots, u_m, v_1, \ldots, v_{n+1}$ дерева t_W (где v_{n+1} – терминальная вершина). Тогда цепочку $b_1 \ldots b_{n-1}$ назовем звеном, если: а) из u_m выходит единственная дуга (u_m, v_1); б) дуга (v_1, v_2) не является единственной, выходящей из v_1, в) если существует вершина v_r $(1 < r < n)$, из которой выходит ровно одна дуга, то путь $(v_r, \ldots, v_{n+1})$ совпадает с линейным участком дерева.

Цепочки e и e' текста T, связанные отношением взаимозамещаемости τ_s, назовем грамматическими единицами, если для каждого из s левых контекстов взаимозамещаемости $\Delta_1, \ldots, \Delta_s$ единиц e и e' имеется сегмент $\Delta_i e_i\#$ со звеном e_i $(1 \leq i \leq s)$. Заметим, что e_i может совпадать, а может и не совпадать с e или e', и e_i может быть, а может и не быть грамматической единицей. Для пояснения рассмотрим дерево t_W из примера I. Пусть $s = 2$. Пути, обведенные сплошной линией, соответствуют цепочкам, являющимся грамматическими единицами. Путь, обведенный пунктирной линией, соответствует цепочке, не являющейся грамматической единицей.

Парадигматическое отношение τ_s – отношение взаимозамещаемости на множестве грамматических единиц.

Синтагматическое отношение S - отношение сочетаемости грамматической единицы с левым контекстом.

Грамматика текста T - множество грамматических единиц с определенными на нем парадигматическими и синтагматическими отношениями.

2. Методы анализа символьных последовательностей, реализованные в ППП СИМВОЛ.

2.1. Реализованные в пакете методы в той или иной мере основываются на 1-граммном анализе символьных последовательностей, т.е. на полном или частичном частотных спектрах 1-грамм, встретившихся в исследуемом тексте.

Получение с помощью существующих ЭВМ таких спектров для текстов большой длины ($N \sim 10^6$ символов) стало реальным после того, как благодаря специально разработанным модификациям методов ассоциативного кодирования (хеширования) удалось снизить трудоемкость соответствующих алгоритмов и сделать ее порядка $l_{max} \cdot N$ [I-4].

В ППП СИМВОЛ используется одна из модификаций этих методов, ориентированная на работу с последовательностями, полностью размещающимися в оперативной памяти, и длиной не более 32000 символов (последнее связано с ограничением на длину символьной строки в версии транслятора с языка ПЛ/I, реализованной в ОС ЕС).

Частотные спектры 1-грамм могут уже сами по себе представлять интерес для исследователей, давая им общую картину о структурных и частотных свойствах интересующих их текстов. Однако основное их назначение - служить исходными данными для других методов исследования.

2.2. В ППП СИМВОЛ реализованы методы представления спектров 1-грамм в виде списков, упорядоченных по частоте встречаемости либо лексикографически, а также в виде деревьев, ветви которых снабжены указаниями частот соответствующих им 1-грамм [5]. Такое представление спектра 1-грамм отличается большой наглядностью и может представлять непосредственный интерес для исследователя (в ППП СИМВОЛ предусмотрен вывод деревьев на печать).

2.3. Методы, ориентированные на выявление общих свойств исследуемых последовательностей, позволяют получать по спектрам 1-грамм такие интегральные характеристики как M_1, E_1^k, l_{max}, F_1^{max}.

С их помощью могут быть получены сложностные оценки символьной последовательности и решены некоторые классификационные задачи (например, по различению "случайных" последовательностей от "неслучайных").

2.4. Методы выделения структурных единиц текста и отношений между ними основываются на анализе спектров 1-грамм, упакованных в виде деревьев [5-6]. Единицы текста выявляются при этом по скачкам частоты и структуре деревьев.

Для текстов на естественных языках флективного типа с помощью этих методов могут выделяться грамматические единицы (основы, флексии и приставки) и далее устанавливаться отношения между ними: парадигматические (отношения взаимозамещаемости между единицами) и синтагматические (определяющие, например, согласованность в употреблении окончаний рядом расположенных слов).

Классификация найденных единиц текста по их парадигматическим и синтагматическим свойствам позволяет установить их грамматическую роль и выявить основные грамматические конструкции, имеющиеся в исследуемом тексте. Таким образом, данные методы позволяют в указанном смысле реконструировать грамматику языка по его текстовой реализации.

Такая реконструкция может представлять интерес не только для анализа неизвестных языков, но и для исследования существующих языков (например, при организации службы языка, оперативно следящей за изменениями как в словарном составе, так и в употребляемых грамматических конструкциях; для определения близости текстов, не только по их словарному составу, но и по употребляемым грамматическим конструкциям, что может использоваться для установления авторства, классификации документов, для выявления неудачных оборотов в системах автоматизации редакционно-издательских работ и т.п.).

Особый интерес может представлять то, что реконструируемая грамматика получается в синтаксической форме, пригодной для непосредственного использования в различного рода алгоритмах дальнейшего анализа текста.

2.5. Особую группу составляют методы совместного анализа нескольких последовательностей с целью определения степени их близости. Это вычисление редакционного расстояния и максимально длинной общей подпоследовательности двух исследуемых последовательностей, определение меры согласованности (близости) нескольких текстов и кластеризация текстов по коэффициенту конкордации.

Эти методы развивались главным образом для исследования генетических последовательностей. Вместе с тем они могут найти применение и для анализа других символьных последовательностей (например, в задаче автоматического обнаружения и коррекции ошибок).

3. Краткое описание элементарных задач, решаемых с помощью ППП СИМВОЛ.

3.1. Вычисление частотного спектра текста. Для заданного текста T длины N ($N \le 32000$ символов) вычисляется частотный спектр (полный или неполный), 1-граммы, входящие в частотные характеристики, упорядочиваются по убыванию частоты встречаемости либо лексикографически.

3.2. Вычисление совместного частотного спектра двух текстов. Для текстов T_1 и T_2, суммарная длина которых не превышает 32000 символов, вычисляется совместный частотный спектр (полный или неполный). 1-граммы, входящие в совместные частотные характеристики, упорядочиваются по убыванию суммарной частоты встречаемости либо лексикографически.

3.3. Вычисление интегральных характеристик. Параметры E_1^k, M_1 и F_1^{max} вычисляются по частотной характеристике 1-го порядка $\Phi_1(T)$, а параметр l_{max} - по полному частотному спектру.

3.4. Генерация текстов. По заданному алфавиту $A=\{a_1,a_2,\ldots,a_n\}$ и вероятностям $P=\{p(a_1),p(a_2),\ldots,p(a_n)/\sum_i p(a_i)=1\}$ генерируется текст T заданной длины N по схеме независимых испытаний.

3.5. Вычисление редакционного расстояния и МДП двух текстов. Для пары текстов T_1 и T_2 длины N_1 и N_2 соответственно вычисляется матрица $D_{N_1 x N_2}$ текущих расстояний, элементы d_{ij} ($1 \le i \le N_1$, $1 \le j \le N_2$) которой есть редакционные расстояния между текстами $T_1[1:i]$ и $T_2[1:j]$. На основе матрицы D определяется редакционное расстояние $d(T_1,T_2)$ между полными текстами и восстанавливается МДП этих текстов. Если матрица D вычислена ранее, она задается для уменьшения трудоемкости наряду с текстами T_1 и T_2 в качестве входного параметра. Восстановление МДП требует в этом случае $O(\min(N_1,N_2))$ операций.

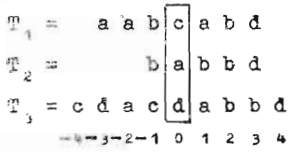
3.6. Определение меры согласованности (близости) $m \ge 2$ текстов в пределах фиксированного окна анализа. Тексты различной длины предварительно располагаются так, чтобы позиции, указанные в векторе фазировки, оказались одна под другой. Например, тексты

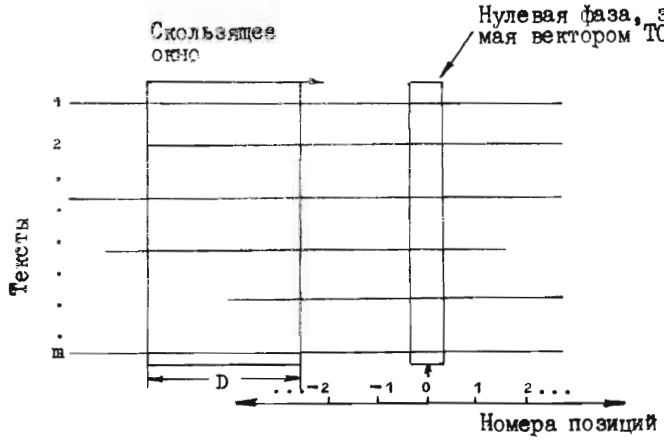
T_1 = a a b c a b d ,

T_2 = b a b b d,

T_3 = c d a c d a b b d

при векторе фазировки TO = (4,2,5) будут выровнены следующим образом:





Фиксируется ширина окна анализа D - число символов, попадающих в рамку (см. рисунок) и задается положение окна (номер позиции, соответствующей крайнему левому символу, охватываемому окном). По последовательностям, не имеющим пустых позиций в пределах окна анализа, вычисляется коэффициент конкордации 1-го порядка, который и характеризует близость текстов в пределах заданного окна анализа.

Окно анализа может скользить с шагом I вдоль текстов слева направо. Для такого режима задаются начальное и конечное положения окна. Значения коэффициента конкордации при каждом положении окна выводятся на график. Сюда же наносятся пороговые значения коэффициента конкордации, соответствующие одно- и пятипроцентному уровням значимости.

3.7. Кластеризация m текстов с коэффициентом конкордации в качестве меры близости. Аналогично п.3.6 тексты фазируются и задается окно для анализа. Последовательности, охватываемые окном, представляются частотными характеристиками порядка 1 . Используя коэффициент конкордации в качестве меры близости группы текстов,

последовательно объединяем (кластеризуем) близкие в указанном смысле тексты. На каждом шаге процесса количество кластеров уменьшается на единицу за счет объединения двух кластеров с максимальным по всем парам значением коэффициента конкордации. Процесс продолжается до тех пор, пока не останется один единственный кластер, либо максимальное значение коэффициента конкордации (по которому выбираются два очередных претендента на слияние) не станет меньше заданного порога кластеризации ε .

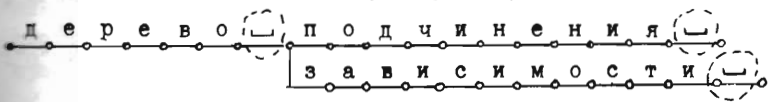
В результате m текстов оказываются разбитыми на группы. При этом коэффициент конкордации текстов, попавших в одну группу, всегда не меньше ε . Варьируя ε , можно получать разбиение на более мелкие либо более крупные группы. Рекомендуемый диапазон ε – (0,5-0,9). Тексты, попавшие в одну группу (кластер), могут быть описаны усредненной частотной характеристикой 1-го порядка.

3.8. Поиск вхождений 1-грамм. Для каждой 1-граммы из заданной группы находятся позиции всех ее вхождений в данный текст.

3.9. Упаковка спектра 1-грамм (или фраз текста) в дерево. Может использоваться для компактного представления спектра (или множества фраз) в памяти машины, обеспечивающего достаточно высокую скорость выборки элементов, а также зачастую является удобной для дальнейшего анализа формой представления данных.

3.10. Подсчет частот встречаемости символов в позициях, наиболее вероятных для разделителей текста. Процессу анализа значимых единиц текста и реконструкции его грамматики должен предшествовать процесс первичной сегментации текста. Один из способов проведения такой сегментации – выделение разделителей текста. Предлагаемый метод позволяет подсчитать частоты встречаемости символов алфавита в позициях, наиболее характерных для разделителей текста (например, непосредственно предшествующих вырождению 1-граммы в единичную и/или непосредственно предшествующих линейному участку дерева, длины, большей 4).

ПРИМЕР 2. Заданный спектр 1-грамм представлен в виде дерева:



Символ "␣" три раза встретился в позиции, наиболее характерной для разделителей текста (эти позиции обведены пунктиром).

3.11. Реконструкция грамматики по тексту (или спектру). Метод реконструкции грамматики опирается на наиболее общие синтаксические свойства грамматических единиц. Перекодировка текста в тер-

минах выделенных единиц позволяет применять метод итеративно. Реконструированная в результате анализа входного текста грамматика состоит из множества грамматических единиц и отношений между ними. Синтагматические отношения выделяются на основе анализа звеньев цепочек W . Структура записи этих отношений: ⟨ левый контекст Δ ⟩ : ⟨ w_1 ⟩, ⟨ w_2 ⟩, ..., ⟨ w_k ⟩, где цепочки w_1 #, w_2 #, ..., w_k # - всевозможные продолжения цепочки Δ в элементах W и хотя бы одна из цепочек w_i ($1 \le i \le k$) является звеном цепочки Δw_i #.

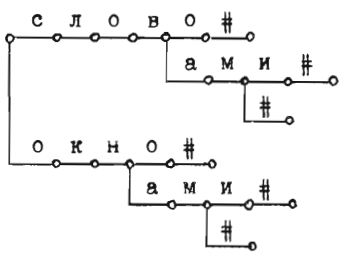
Синтагматическое отношение связывает цепочки Δ и w_i для $1 \le i \le k$. Парадигматическое отношение понимается как отношение взаимозамещаемости. Структура записи выделенных парадигматических отношений:

$$\langle e_1 \rangle: \quad \langle e_{11} \rangle = \langle e_{12} \rangle = \langle e_{13} \rangle = \ldots = \langle e_{1n} \rangle ,$$

$$\langle e_2 \rangle: \quad \langle e_{21} \rangle = \langle e_{22} \rangle = \langle e_{23} \rangle = \ldots = \langle e_{2m} \rangle ,$$

где e_i, e_{ij} - грамматические единицы; парадигматическое отношение связывает любую пару единиц e_i и e_{ij} . Грамматическими единицами считаются элементы области определения парадигматических отношений.

ПРИМЕР 3. Заданное множество сегментов представлено в виде дерева:



Реконструированная грамматика будет содержать:

а) синтагматические отношения: "слов": "о", "ами", "ам"; "окн": "о", "ами", "ам";

б) парадигматические отношения: "ам": "ами" = "о"; "о": "ами" = "ам"; "ами": "о" = "ам";

в) грамматические единицы: "о", "ами", "ам".

4. Возможные сферы использования ППП "СИМВОЛ".

4.1. Построение машинных словарей с автоматическим объединением словоформ, относящихся к одной лексеме. Словоформы объединяются на основе общности левого контекста при синтагматических от-

ношениях. Задача возникает в информатике, при автоматическом переводе, автоматизации лингвистических исследований.

4.2. Создание службы языка, объективно следящей за происходящими в нем изменениями и позволяющей своевременно вносить их в формальную модель грамматики.

4.3. Автоматическое обнаружение и коррекция ошибок в тексте. Анализ частотного спектра позволяет выявлять отдельные типы ошибок в тексте. Укажем, для примера, на следующие возможности. При малых значениях l и достаточно большой длине текста N $(N/n^l >> 1)$ мы вправе ожидать достаточно высокой частоты встречаемости в тексте каждой из разрешенных (в данном языке) l-грамм. Поэтому наличие отличных от нуля значений параметров E_l^k при малых l и k чаще всего сигнализирует об ошибках типа "появление запретной комбинации" ("оы", "ъы", "чс","еь").

Если представить множество l-грамм текста в виде дерева, то информация о возможных ошибках может быть почерпнута из анализа характера ветвления этого дерева. Разветвления с аномально низкой частотой зачастую сигнализируют об ошибке. Например, если I9-грамма "␣ электронных ␣ цифров" встречается в тексте I4 раз, а 20-грамма "␣ электронных ␣ цифровы" I3 раз,то ответвление с единичной частотой с большой вероятностью соответствует ошибке.

Расположение l-грамм в тексте также может нести информацию об ошибке. Так, если в пределах небольшого участка текста встречаются две похожие (например, в смысле максимально длинной общей подпоследовательности) l-граммы ("статический - статистический", "наибольшие - небольшие","форматный - формантный"), то одна из них (как правило, та, что имеет меньшую частоту) могла появиться как ошибка из-за похожести написания или звучания.

Процедуры типа вычисления редакционного расстояния могут использоваться не только для обнаружения ошибок, но и для исправления их. Коррекция ошибок, как правило, основывается на поиске в словаре слов, близких в определенном смысле к тому, в котором допущена ошибка.

4.4. Сжатие текстов. Для сжатия текстов без потери имеющейся в них информации часто используют неравноблочные коды (Хаффмена, Шеннона-Фано и др.). Идея сжатия заключается в присвоении наиболее часто встречающимся символам (или группам символов) наиболее коротких кодов. Информация, необходимая для построения таких кодов, содержится в полном частотном спектре текста.

4.5. Выявление структурных свойств текстов заданного типа. Как показали исследования, реконструированные грамматики текстов разных профессиональных или авторских стилей различаются как по множествам грамматических единиц и отношений, так и по характеру отклонений от традиционной грамматики. Задача возникает при анализе и сопоставлении профессиональных или авторских стилей; при выработке ограничений на использование грамматических средств в системах автоматического редактирования; при автоматическом генерировании предметно-ориентированных машинных моделей грамматик в информатике и автоматическом переводе; при анализе текстов неизвестных языков.

4.6. Классификация текстов.

4.6.I. В экспериментах с текстами на естественном языке и генетическими текстами (первичными структурами нуклеотидных молекул простейших микроорганизмов) прослеживалась одна общая закономерность: чем грубее классификация, тем меньшего порядка частотные характеристики используются в качестве классификационных признаков.

Так, при обработке текстов на естественном языке (художественном, политическом, техническом, $N_i \sim I0^5$ символами, $1 \leq i \leq 3$) общеязыковые закономерности просматривались лишь для малых значений i ($i = 2,3$). При бо́льших значениях i начинает уже сказываться специфика текста (художественный, политический, технический), его содержание, индивидуальность автора и т.д.

К примеру, составы первого (по частоте встречаемости) десятка би- и триграмм по каждому из трех видов текста сильно коррелированы (закономерность, характеризующая язык в целом), однако порядок этих 1-грамм (их ранги) уже отражает специфику текста. Так, в техническом тексте наиболее частой является биграмма "ен" (выражение, отношение, заряженный и т.д.), которая в художественном и общественно-политическом текстах стоит соответственно на 38-м и I2-м местах; в общественно-политическом тексте на первом месте стоит биграмма "ст" (капиталистический, империалистический и т.д.), занимающая соответственно I5-е и 2-е место в художественном и техническом текстах.

Конкретное содержание текста является более частной характеристикой по сравнению с его тематической направленностью, поэтому для его выявления требуются 1-граммы большей длины. К примеру, на 5-м месте по частоте встречаемости среди 5-грамм в художественном

тексте стоит "Иван␣", который является главным действующим лицом в одном из отрывков этого текста. Наиболее частыми IO-граммами технического текста являются "поверхност", "␣потенциал", "электричес" и т.д., которые свидетельствуют о том, что в данный текст вошли статьи по электронике.

Еще более частной характеристикой является индивидуальный стиль автора, для выявления которого требуются соответственно еще более длинные 1-граммы. Так, среди наиболее частых 2I-грамм фигурируют "сказал Иван Барабанов" (9 раз), "сказал Иван Фролов,␣" (8 раз), "сказал Иван Алеков,␣" (6 раз), которые, вероятнее всего, отражают индивидуальный почерк автора.

Аналогичная картина наблюдается и для генетических текстов ($N \sim 5 \cdot 10^3$ символов, $n = 4$). Так для классификации микроорганизмов на два класса: "эукариоты и прокариоты" по их первичным структурам (текстам с указанными выше параметрами), в принципе, достаточно знания частотной характеристики второго порядка. Биграмма "CG" в "эукариотических текстах" имеет аномально низкую встречаемость по сравнению с остальными биграммами, чего не наблюдается у прокариотов. Для более детальной классификации отдельно прокариотов и эукариотов требуется привлечение 1-граммных характеристик более высокого порядка.

4.6.2. Обычно не все участки текстов являются информативными в плане интересующей нас задачи классификации. Примером могут служить знаки пунктуации в генетических текстах. Это короткие последовательности длиной от трех до нескольких десятков символов, ответственные за начало (или окончание) основных генетических процессов (редупликации, транскрипции, трансляции). Типичная схема организации таких знаков пунктуации - наличие нескольких (2-3) информативных участков, разделенных группами неинформативных символов.

Если имеется обучающая выборка одинаково сфазированных знаков пунктуации одного вида, то положение информативных участков, их размер, а иногда и значимость могут определяться с помощью анализа обучающей выборки скользящим окном переменной длины. Критерием информативности участка при этом выступает значение коэффициента конкордации частотных характеристик 1-го порядка (для всех последовательностей обучающей выборки) в пределах заданного окна.

4.6.3. Объекты обучающей выборки даже в пределах одного класса могут быть достаточно разнородны. Для построения решающего

правила часто бывает целесообразно произвести предварительное разбиение обучающей выборки на группы относительно однородных объектов. Подобное разбиение может быть выполнено при помощи алгоритма кластеризации символьных последовательностей с коэффициентом конкордации в качестве меры близости объектов внутри группы.

Приемы, описанные в п.4.6.2 и 4.6.3, использовались при построении алгоритмов обнаружения знаков пунктуации в генетических текстах.

4.7. Выявление эволюционной близости генетических текстов. Эволюция генетических текстов может быть описана элементарными преобразованиями вида:

а) замена одного символа другим;

б) вставка или устранение группы символов;

в) дупликация группы символов;

г) инверсия группы символов (преобразование, сводимое при помощи перекодировки к дупликации);

д) транспозиция (перенос) группы символов из одного места в другое.

Часто возникает задача установления соответствия (гомологии) двух текстов, описывающих микроорганизмы, разошедшиеся в процессе эволюции. Она сводится к установлению наиболее правдоподобного (осмысленного с молекулярно-биологической точки зрения) числа и типа элементарных операций, которыми один текст может быть переведен в другой.

Алгоритм отыскания МДП неприменим в чистом виде из-за того, что большинство элементарных операций носит "групповой" характер (т.е. затрагивает группу символов). Наиболее естественный путь сводится в комбинации двух приемов: отысканию "точек синхронизации" (фазировки) двух текстов в виде повторов или инверсий достаточно большой длины и отбрасыванию "ложных" точек фазировки, возникающих в силу случайных обстоятельств. Первый прием может быть реализован с помощью алгоритма получения совместного частотного спектра, а второй - при помощи алгоритма отыскания МДП среди претендентов, полученных на первом этапе и упорядоченных в соответствии с адресами их вхождений в свои тексты (отыскание МДП в новом "укрупненном" алфавите). Итеративное повторение этих двух шагов (на все более мелких участках) позволяет осуществлять все более детальное сопоставление пар текстов.

4.8. Проверка последовательностей на "случайность". Для чисто случайных последовательностей могут быть получены математические и

дисперсии многих из введенных выше параметров описания, например, таких как M_1, E_1^k, l_{max} и т.д. Имея аналитические выражения для различных моментов этих случайных величин и получая на основе вычисления частотного спектра выборочные оценки указанных моментов, можно строить различные тесты для проверки последовательностей на случайность, используя критерии согласия.

4.9. Редактирование текстов. При редактировании текста предполагается, что все входящие в него символы занумерованы, начиная с единицы. Для выделения цепочки текста используется конструкция ИДЕР (N1,N2), где ИДЕР - имя редактируемого текста, N1 - номер символа от начала текста, с которого начинается цепочка, N2 - номер символа, которым кончается цепочка. Необходимо, чтобы $N1 \leq N2$. Если N1 = N2 или N2 отсутствует, то выделяется цепочка длиной в один символ с номером N1 . Этот способ выделения цепочек положен в основу следующих редактирующих операций: вставки перед выделенной цепочкой; вставки после выделенной цепочки; замены выделенной цепочки; удаления выделенной цепочки.

К операциям редактирования относятся также операции переко - дирования текстов. При этом необходимо задавать соответствие символов исходного алфавита символам нового алфавита.

5. Организация ППП СИМВОЛ

ППП СИМВОЛ включает следующие компоненты: функциональное наполнение, отражающее конкретный класс задач обработки символьных последовательностей; системное наполнение, автоматизирующее работу с пакетом; язык общения с пакетом. Функциональное наполнение пакета включает обрабатывающие модули, модули обеспечения интер - фейсов и связей, написанные на языках ПЛ/I, Р/ТРАН, Ассемблер и оформленные в виде подпрограмм с параметрами. Обращение к модулям осуществляется неявно через определенные конструкции языка пакета.

Системное наполнение пакета включает следующие компоненты: транслятор с входного языка пакета; управляющую программу; сер - висные программы.

Транслятор с языка пользователя проверяет синтаксис задания и переводит его в некоторый внутренний язык.

Управляющая программа осуществляет ввод и трансляцию зада - ний пользователя; составление последовательности модулей, реали - зующих задание; интерпретацию задания пользователя с внутреннего языка; загрузку и передачу управления всем компонентам пакета;

взаимодействие с наборами данных, находящихся на ленте или диске; контроль на полноту входных данных перед запуском отрабатывающих модулей.

Сервисные программы предназначены для ведения библиотечного хозяйства пользователя (ввода, коррекции и распечатки данных). По отношению к аналогичным средствам операционной системы сервисные программы пакета более просты и доступны массовому пользователю, хотя имеют ограниченные возможности.

ППП СИМВОЛ реализован погружением в операционную систему, т.е. максимально использует ее средства и не накладывает ограничений на использование этих средств в рамках пакета. Это обеспечивает следующие функциональные характеристики ППП СИМВОЛ:

задание на работу с пакетом может появиться наряду с другими заданиями для ОС ЕС и, с другой стороны, все средства языка управления заданиями ОС ЕС могут использоваться в пакете без ограничений;

сборка программы из модулей, хранящихся в библиотеках пакета, может осуществляться средствами ОС ЕС;

для обслуживания входных и выходных данных (размещаемых в обычных наборах данных ОС ЕС), а также их корректировки могут применяться сервисные программы (утилиты) ОС ЕС.

6. Язык и схема работы с ППП СИМВОЛ

Входной язык является средством общения с пакетом и предназначен для записи задания на обработку символьных последовательностей. Средства языка позволяют не задавать явно алгоритмов решения, а ограничиться лишь спецификациями входных и выходных данных (т.е. формализованными описаниями характеристик данных).

Задание на входном языке состоит из двух частей: спецификации входов - описания носителя, на котором располагаются входные данные. и их характеристик, а также спецификации выходов, описывающих результат. Термины и конструкции входного языка отражают специфику предметной области.

Для обращения к ППП СИМВОЛ пользователю необходимо сформулировать свое задание на языке пакета - описать спецификации входных и выходных данных.

Кроме языка пакета, пользователь использует язык управления заданиями ОС ЕС в упрощенном виде. С его помощью пользователь уп-

равляет разными видами работ. Для каждого из этих видов разработаны каталогизированные процедуры, организованные в библиотеку каталогизированных процедур. Каталогизированные процедуры - это наиболее часто употребимые при работе с пакетом предложения языка управления заданиями. Квалифицированный программист может расширить эту библиотеку своими процедурами обычным образом.

Пример задания и результаты счета приведены на распечатке. Ключевые слова языка пользователя подчеркнуты.

Литература

I. Ассоциативное кодирование: реализация и применение /Величко В.М., Гусев В.Д., Косарев Ю.Г., Лозовский В.С., Титкова Т.Н. -В кн.: Вычислительные системы. Вып. 62. Ассоциативное кодирование. Новосибирск, 1975, с.3-37.

2. ГУСЕВ В.Д., КОСАРЕВ Ю.Г., ТИТКОВА Т.Н. О задаче поиска повторяющихся отрезков текста. - Там же, с. 49-71.

3. ГУСЕВ В.Д., КОСАРЕВ Ю.Г., ТИТКОВА Т.Н. Отыскание статистических закономерностей текстов методом ассоциативного кодирования. -Там же, с. 72-89.

4. ГУСЕВ В.Д., КОСАРЕВ Ю.Г., ТИТКОВА Т.Н. Методы поиска и анализ статистических закономерностей в символьных последовательностях. -В кн.: Машинные методы обнаружения закономерностей. (Материалы Всесоюз. симпозиума, 5-7 апреля 1976 г.). Новосибирск, 1976, с. 75-84.

5. ТИМОФЕЕВА М.К. Применение Р-технологии программирования для организации больших словарей в памяти ЭВМ. -В кн.: Автоматизированные системы управления ВУЗом. Новосибирск, НГУ, 1978, с.57-66.

6. ТИМОФЕЕВА М.К. Индуктивная реконструкция грамматик флективных языков. -В кн.: Методы обнаружения закономерностей с помощью ЭВМ (Вычислительные системы, вып. 91). Новосибирск, 1981, с. 57-67.

Поступила в ред.-изд.отд.
24 октября 1983 года

```
//  EXEC  SYMBOL
//LKED.SYSIN DD *
IEF142I - STEP WAS EXECUTED - COND CODE 0000
IEF373I STEP /LKED     / START 83307.0858
IEF374I STEP /LKED     / STOP  83307.0900 CPU    0MIN 29.52SEC MAIN 204K
//SYM.SYSIN  DD   *

ЗАДАНИЕ  PARAD
ВХОД
РАЗДЕЛИТЕЛИ RAZD(5)=' ,.;!'
ДЕРЕВО TANGL ИЗ БИБЛИОТЕКИ TREELIB
ВЫХОД
ПАРАДИГМАТИЧЕСКИЕ ХАРАКТЕРИСТИКИ ДЛЯ S=1 ПЕЧАТЬ В 1 ЭКЗ.
КОНЕЦ
***СТАТИСТИКА***
***СИНТАКСИЧЕСКИХ ОШИБОК НЕТ***
***СЕМАНТИЧЕСКИХ ОШИБОК НЕТ***


*****
РАЗБИЕНИЯ ВХОДНЫХ ЦЕПОЧЕК
*****

PETITION;@;S;
PERMIT;@;S;TING;
PERSONAL;LY;@;
PERSONIALLY:AL;S;
PREVIOUS;@;LY;
PROPOSIED;AL;
PROVIDE;D;@;
PROMISE;D;@;
REPUBLIC;ANS;@;
REPORTIEDLY;@;
REPRESENT;ING;ATIVES;
RESIGN;ED;ATION;
RESOLUTION;@;<;
REQUISESTING;IRE;IREMENT;
REQUIRE;@;MENT;
RECONSITRUCTION;IDERATION;
RECOMMEND;@;ATIONS;
REJECT;ED;ION;
ROBERT;@;S;
ROAD;S;@;
SIGN;TURES;@;
SITE;@;S;
SENATIE;ORS;OR;
SENATOR;S;@;
SESSION;@;S;
SEVE;N;RL;
STUD;Y;ENTS;
SAVINGS;@;
SPECI;FICALLY;AL;
SHER;IFF;MAN;
SCHOOL;@;S;ING;
SCHOLASTIC;@;S;
GEORG;IA;E;
GIVE;N;@;
GIFT;@;S;


*****
ГРАММАТИЧЕСКИЕ ЕДИНИЦЫ И ОТНОШЕНИЯ:
*****

@:  S=LY=D=ING=N=
S:  AL=ING=@=
LY:  @=
AL:  ED=S=
ED:  AL=
D:  @=
ING:  @=S=
N:  @=

IEF142I - STEP WAS EXECUTED - COND CODE 0000
IEF373I STEP /SYM      / START 83307.0900
IEF374I STEP /SYM      / STOP  83307.0902 CPU    0MIN 44.32SEC MAIN 206K
```

```
//SYMBOL JOB '12 ЛАБ','ЧУЖАНОВА',MSGLEVEL=(2,0),REGION=250K
//  EXEC SYMBOL
//LKED.SYSIN DD *
 EF142I - STEP WAS EXECUTED - COND CODE 0000
IEF373I STEP /LKED     / START 83307.0856
IEF374I STEP /LKED     / STOP  83307.0858 CPU    0MIN 35.44SEC MEIN 204K
//SYM.SYSIN DD *

ЗАДАНИЕ SPECTR2
ВХОД
АЛФАВИТ AL(4)='АГТЦ'
ТЕКСТЫ T1(25)='ААЦЦЦГГГГГТАЦГТАЦГАГЦАГЦГГ',T2(20)='АГГТГГЦЦААГЦЦА
ГТГАЦЦГ'
ВЫХОД
СПЕКТР ДЛЯ ПАРЫ ТЕКСТОВ,УПОРЯДОЧЕННЫЙ ПО АЛФАВИТУ,ДЛЯ L=2
ПЕЧАТЬ В 1 ЭКЗ.
КОНЕЦ
***СТАТИСТИКА***
***СИНТАКСИЧЕСКИХ ОШИБОК НЕТ***
***СЕМАНТИЧЕСКИХ ОШИБОК НЕТ***



ПАРАМЕТРЫ:              L=2                    ML=EL1=12              EL1=2                  MAX=6


|          |          |          |          |
|----------|----------|----------|----------|
| L=2      | T1       | T2       | T1&T2    |
| ML-EL1   | 1        | 1        | 10       |
| EL1      | 0        | 2        |          |


АЛФАВИТНОЕ УПОРЯДОЧЕНИЕ                    * * *  С Т А Т И С Т И К И  2-ГРАММ
 * * *
АА-1+1    АЦ-3+1    ГГ-4+2    ГЦ-2+2    ТГ-0+2    ЦА-1+1    ЦГ-4+2    ЦЦ-2+3
АГ-2+2    ГА-1+1    ГТ-2+2    ТА-2+0


EF142I - STEP WAS EXECUTED - COND CODE 0000
IEF373I STEP /SYM      / START 83307.0858
IEF374I STEP /SYM      / STOP  83307.0858 CPU    0MIN 03.26SEC MEIN 198K
```