МАШИННЫЕ МЕТОДЫ ОБНАРУЖЕНИЯ ЗАКОНОМЕРНОСТЕЙ
(Вычислительные системы)
1981 год Выпуск 88

УДК 519.766.4

ХАРАКТЕРИСТИКИ СИМВОЛЬНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

В.Д.Гусев

Описан способ представления символьных последовательностей, ориентированный на решение многих содержательных задач из области лингвистики, генетики, теории связи. В качестве примера можно указать на задачи автоматического обнаружения и коррекции ошибок в тексте, сжатия текста без потери информации, вычисления эволюционных расстояний (между белками, геномами), анализа первичной структуры нуклеотидных молекул с целью автоматического выявления участков с различным функциональным назначением и ряд других.

Проблема представления играет важную роль в задачах распознавания образов. Выбор удачного описания (информативной системы признаков), как правило, позволяет избежать использования сложных решающих правил и повышает надежность распознавания.

Изложение сопровождается примерами, иллюстрирующими результаты обработки по нижеописанной схеме текстов на естественном языке и генетических текстов. Последние были представлены первичными структурами некоторых полностью расшифрованных молекул ДНК простейших микроорганизмов - вирусов и бактериофагов. Длины таких текстов порядка нескольких тысяч символов, а алфавит состоит из четырех символов (нуклеотиды А,Т,Г,Ц).

§1. Специфика "символьных" объектов

Объекты, описываемые символьными последовательностями, несколько выпадают из традиционной для распознавания схемы представления исходных данных (в виде таблицы "объект-свойство") и ориентированной на нее схемы анализа. Это объясняется следующими обстоятельствами.

I. В отличие от табличного представления, где порядок признаков, как правило, несуществен, символы в последовательности упорядочены, и информация о порядке играет существенную роль при классификации "символьных" объектов.

2. Многие объекты, описываемые символьными последовательностями, эволюционируют во времени либо искажаются под действием помехи. Эволюция (искажение) объектов в большинстве случаев может быть описана следующими преобразованиями: а) устранением символа (или группы символов), б) добавлением символа (или группы символов), в) заменой одного символа другим.

Интегральные характеристики символьных последовательностей не должны быть слишком чувствительными к преобразованиям такого рода.

3. Каждый символ можно интерпретировать как результат измерения в "слабой" шкале - шкале наименований. Бедность шкалы накладывает свой отпечаток и на модели, используемые для описания символьных последовательностей. Используются модели "неарифметической" природы, такие, например, как порождающие грамматики или марковские цепи соответствующего порядка.

Применение моделей первого вида затруднено из-за отсутствия в общем случае формальных процедур восстановления грамматики по обучающей выборке (именно такая ситуация представляет интерес в задачах распознавания образов). Для построения марковских моделей требуются оценки переходных вероятностей, т.е. условных вероятностей перехода от произвольной l-граммы (связной подпоследовательности, состоящей из l символов) $a_{i_1} \ldots a_{i_l}$ к произвольному символу $a_{i_{l+1}}$ алфавита A ($a_{i_k} \in A$, $1 \le k \le l+1$, $1 \le i_k \le n$, где $n = |A|$). Чтобы оценки были достоверными, необходимы выборки очень большого объема даже для небольших значений l. Получение статистик l-грамм становится при этом нетривиальной вычислительной задачей.

Следует отметить также, что во многих приложениях приходится иметь дело с относительно короткими последовательностями, не обязательно вероятностной природы, представленными к тому же в ограниченном количестве. Применение марковских моделей в подобной ситуации некорректно. Представляется, что в этом случае полезным может оказаться описание, выявляющее периодическую структуру последовательности. Выявление и подсчет всех повторов, содержащихся в тексте, назовем l-грамммным анализом. Результатом его является представление объекта в соответствии с тради -

ционной схемой "объект-свойство", для которой в рамках распознавания образов разработан достаточно богатый аппарат анализа.

§2. Методика 1-граммного анализа текста

I. Частотные (первичные) характеристики текста. Всякий текст

$$T = a_1 a_2 a_3 a_4 a_5 \ldots a_N \qquad (1)$$

длины N над алфавитом A мощности n может быть представлен в виде последовательности частично перекрывающихся (в общем случае) 1-грамм (l = I,2,..., N), каждая из которых выделяется путем сдвига на один символ вдоль текста скользящего окна длиной в l символов. Для примера в последовательности (I) выделены подчеркиванием первые три 3-граммы текста. Полное число 1-грамм, содержащихся в тексте, равно N - l+1. Поскольку среди них могут быть повторяющиеся, число различных 1-грамм $M_l \le N - l + 1$.

Подсчитывая число повторений каждой из разновидностей 1-грамм в тексте, получаем частотную характеристику текста порядка 1: $\Phi_l(T) = \{\varphi_{l1}, \varphi_{l2}, \ldots, \varphi_{lM_l}\}$. Каждый элемент φ_{lr} $(1 \le r \le M_l)$ частотной характеристики есть пара " r-я 1-грамма - частота $F_l(r)$ ее встречаемости в тексте". Формы представления частотных характеристик могут быть разными.

Так, например, при решении задачи оптимального кодирования текста с целью его сжатия целесообразно упорядочить все 1-граммы по убыванию частоты встречаемости $F_l(r)$ $(F_l(1) \ge F_l(2) \ge \ldots \ge F_l(M_l))$. Параметр r определяет в этом случае ранг соответствующей 1-граммы. При наличии статистической избыточности в тексте его сжатие осуществляется путем присвоения наиболее часто встречающимся 1-граммам наиболее коротких кодов. Данный способ представления частотных характеристик используется также для вычисления ранговых мер сходства двух символьных последовательностей.

Если возникает необходимость в многократном поиске нужных 1-грамм (например, при вычислении некоторых 1-граммных мер сходства двух символьных последовательностей), то частотная характеристика текста может быть представлена в виде хэш-таблицы [I], что, как

правило, гарантирует быстрый поиск каждой l-граммы. При этом, однако, l-граммы внутри хэш-таблицы уже не упорядочены по частоте.

И наконец, если нас интересует дерево всевозможных расширений при переходе от l- к l+1-граммам (l = 1,2,...), все частотные характеристики удобно представить в виде лексикографических деревьев (каждый символ алфавита порождает одно дерево) [2]. При этом весьма наглядной оказывается структура текста, выявляются его стабильные и неустойчивые элементы ("корни и флексии"), однако время поиска l-грамм увеличивается по сравнению с предыдущим случаем и разрушается упорядоченность l-грамм по частоте.

Интуитивно ясно, что в частотной характеристике l-го порядка содержится больше информации, чем в характеристике (l-1)-го порядка. Рассмотрим возможность восстановления $\Phi_{l-1}(T)$ из $\Phi_l(T)$.

Каждая l-грамма $a_{i_1} a_{i_2} \ldots a_{i_{l-1}} a_{i_l}$ частотной характеристики содержит две (l-1)-граммы: левостороннюю $a_{i_1} a_{i_2} \ldots a_{i_{l-1}}$ и правосторонюю $a_{i_2} \ldots a_{i_{l-1}} a_{i_l}$. Разобьем множество l-грамм, входящих в $\Phi_l(T)$, на классы эквивалентности $\hat{S}^{Л}_1, \ldots, \hat{S}^{Л}_{M_Л}$, состоящие каждый из l-грамм с совпадающими левосторонними (l-1)-граммами, и вычислим значения $F^{Л}_{l-1}(r) = \sum_{i \in S^{Л}_r} F_l(i)$, $1 \le r \le M_Л$. Аналогичное разбиение $(S^{П}_1, \ldots, S^{П}_{M_П})$ проведем по правосторонним (l-1)-граммам и вычислим $F^{П}_{l-1}(r) = \sum_{i \in S^{П}_r} F_l(i)$, $1 \le r \le M_П$. Классам эквивалентности, соответствующим одной и той же (l-1)-грамме в обоих разбиениях, присвоим одинаковые номера.

Поскольку каждая (l-1)-грамма текста (за исключением начальной $r_{нач}$ и конечной $r_{кон}$) является одновременно левосторонней (для одной из охватывающих ее l-грамм) и правосторонней (для другой), она учитывается по разу в обоих разбиениях. Поэтому $F^{Л}_{l-1}(r) = F^{П}_{l-1}(r) = F_{l-1}(r)$ для всех $r \ne r_{нач}$ и $r \ne r_{кон}$. Начальная (l-1)-грамма текста может быть лишь левосторонней и учитывается только в первом разбиении, конечная - только правосторонней и учитывается только во втором. Следовательно, если начальная и конечная (l-1)-граммы н е с о в п а д а ю т, то $F^{Л}_{l-1}(r_{нач}) - F^{П}_{l-1}(r_{нач}) = 1$ и $F^{П}_{l-1}(r_{кон}) - F^{Л}_{l-1}(r_{кон}) = 1$. Эти соотношения позволяют идентифицировать начальную и конечную (l-1)-граммы текста и доопределить $\Phi_{l-1}(T)$: $F_{l-1}(r_{нач}) = F^{Л}_{l-1}(r_{нач})$ и

$F_{l-1}(r_{кон}) = F^{П}_{l-1}(r_{кон})$. Таким образом, в этом случае $\Phi_{l-1}(T)$ однозначно восстанавливается по $\Phi_l(T)$.

Если начальная и конечная (l-1)-граммы с о в п а д а ю т ($r_{нач} = r_{кон} = r_{гран}$), то $F^{Л}_{l-1}(r_{гран}) = F^{П}_{l-1}(r_{гран}) = F_{l-1}(r_{гран}) - 1$, т.е. мы не можем идентифицировать их среди остальных (l-1)-грамм по различию частот в обоих разбиениях, а лишь можем утверждать, что частота одной из (l-1)-грамм занижена на единицу. Таким образом, в данном случае $\Phi_{l-1}(T)$ однозначно не восстанавливается по $\Phi_l(T)$.

Определим параметр l_{max} как минимальное l , начиная с которого $F_l(r) \equiv 1$ для всех r , т.е. в тексте уже отсутствуют повторяющиеся l-граммы. Совокупность **частотных характеристик** для значений $l = 1,2,\ldots,l_{max}+1$ назовем п о л н ы м ч а с т о т н ы м с п е к т р о м текста. Эпитет "полный" означает, что по частотной характеристике порядка $(l_{max}+1)$ исходный текст уже восстанавливается однозначно.

Действительно, для восстановления текста достаточно определить его начальную (или конечную) $(l_{max}+1)$-грамму и воспользоваться перекрываемостью каждой пары соседних $(l_{max}+1)$-грамм на участке длиной в l_{max} символов. Поскольку из определения l_{max} следует, что $F_{l_{max}}(r) \equiv 1$ для всех r , каждый раз будет находиться е д и н с т в е н н а я $(l_{max}+1)$-грамма, имеющая общее перекрытие с предыдущей на участке из l_{max} символов. Определение начальной (или конечной) $(l_{max}+1)$-граммы осуществляется аналогично описанному выше. Неоднозначности здесь не возникает, поскольку начальная и конечная (l-1)-граммы (в данном случае это l_{max}-граммы) не совпадают.

Заметим, что по частотной характеристике порядка l_{max} в общем случае уже невозможно однозначно восстановить исходный текст. В качестве примера приведем две последовательности: T_1 = = abcabdaba и T_2 = abdabcaba с идентичными частотными характеристиками в диапазоне от $l = 1$ до $l = l_{max} = 3$, но отличающиеся порядком следования элементов.

Интересно указать на содержательный аналог задачи восстановления последовательности по фрагментам, возникающий при определении первичной структуры ДНК- и РНК-молекул. С помощью существующих методик удается восстанавливать первичную структуру лишь относительно коротких участков (~20-300 нуклеотидов). Поэтому для определения первичной структуры всего генома его разрезают на более короткие перекрывающиеся участки специальными ферментами-рест-

риктазами, определяют первичную структуру каждого участка, а затем, выявляя перекрытия, восстанавливают геном в целом. Оценки параметра l_{max} (вероятностные, экспериментальные) определяют минимально необходимые для однозначного восстановления длины перекрытий фрагментов.

2. Интегральные (вторичные) характеристики текста. Параметры, вычисляемые на основе частотных (или первичных) характеристик текста, назовем интегральными (или вторичными) характеристиками текста. Таковыми, например, являются введенные выше параметры M_l и l_{max}. Укажем ряд других интегральных характеристик, играющих важную роль при классификации текстов.

2.1. Пусть E_l^k - количество различных l-грамм, каждая из которых встречается в тексте ровно k раз ($k = 1,2,\ldots,N-l+1$) или не встречается ни разу ($k = 0$). Совокупность значений E_l^k, упорядоченная по возрастанию параметра k (при фиксированном l), является аналогом известной в лингвистике кривой Юла, отражающей в некотором смысле богатство и индивидуальные особенности языка автора. Параметры N, M_l, n связаны с совокупностью характеристик E_l^k следующими простыми соотношениями (предполагается, что $F_l(1) \geq F_l(2) \geq \ldots \geq F_l(M_l)$):

$$M_l = \sum_{k=1}^{F_l(1)} E_l^k, \qquad (2)$$

$$N-l+1 = \sum_{k=1}^{F_l(1)} k \cdot E_l^k, \qquad (3)$$

$$E_l^0 = n^l - M_l. \qquad (4)$$

Параметр E_l^0 характеризует количество запрещенных комбинаций, или разрешенных, но не встретившихся в данном конкретном тексте. Параметры E_l^1, E_l^2 при больших N и малых l зачастую отличны от нуля лишь вследствие появления запретных комбинаций, возникающих в результате ошибок при подготовке данных. Значение $l_{max}-1$ соответствует самому длинному повтору в тексте. Как правило, такие повторы являются функционально значимыми, особенно, когда расположены рядом (например, дупликации в генетических текстах). Значение l^*, при котором достигается $\max_l(M_l - E_l^1)$, соответствует пику

трудоемкости алгоритма, последовательно вычисляющего $\Phi_1(T)$. Интересно отметить, что несмотря на большие различия между текстами на естественном языке и генетическими как по размеру алфавита ($n_1 = 35$, $n_2 = 4$), так и по длине ($N_1 \sim 10^5$-$5 \cdot 10^5$, $N_2 \sim 5 \cdot 10^3$), значения 1^* практически совпадали ($1^* \sim 6$-8).

2.2. Для компактного описания частотных характеристик, представленных в упорядоченной форме, можно использовать различного рода аппроксимации, например, вида

$$F_1(r) = Nk_1(r+r_1)^{-\gamma_1}, \quad k_1, \gamma_1 > 0, \qquad (5)$$

где k_1, r_1, γ_1 - параметры, подлежащие оцениванию. Зависимость (5) является аналогом известной формулы Ципфа-Мандельброта [3], используемой для описания закона убывания частот встречаемости слов в частотных словарях.

Наибольший интерес представляет параметр γ_1, характеризующий средний наклон частотной характеристики. Оценки этого параметра на текстах естественного языка (слова, биграммы, триграммы) дают значения, близкие к единице. Оценки γ_2 и γ_3 для генетических текстов заметно варьируют, но, как правило, не превышают 0,5.

2.3. Для описания достаточно длинных текстов могут быть использованы понятия энтропии и избыточности. Шеннон [4] определил условную энтропию H_1 порядка (1-1), т.е. энтропию, отвечающую опыту по выявлению 1-й буквы текста при наличии информации о предыдущих (1-1)буквах, в виде

$$H_1 = -\sum_{i,j} P(b_i,j) \log_2 Pb_i(j), \qquad (6)$$

где b_i - блок из (1-1)-й буквы ((1-1)-грамма); j - произвольная буква, следующая за b_i; $P(b_i,j)$ - вероятность встречаемости 1-граммы b_ij; $Pb_i(j)$ - условная вероятность следования буквы j за блоком b_i. Соответственно избыточность (1-1)-го порядка имеет вид

$$r_1 = 1 - \frac{H_1}{\log_2 n}. \qquad (7)$$

Последняя характеристика более удобна для классификации, поскольку нормирована к диапазону 0-1.

Следует отметить, однако, что с увеличением l (при фиксированном N) оценки вероятностей, используемые для вычисления (6), становятся статистически недостоверными, т.е. мы можем пользоваться характеристиками (6) и (7) лишь для небольших значений l .

Естественно-языковые и генетические тексты резко отличаются друг от друга по значениям параметра R_1. Оценки R_4, к примеру, для естественного языка были близки к 0,6 ($N = 5 \cdot 10^5$, $n = 35$); для генетических текстов $\hat{R}_4 \sim 0{,}05$–$0{,}07$ ($N \sim 5 \cdot 10^3$, $n = 4$) .

§3. Вычисление частотных характеристик

Можно отметить три подхода, в той или иной степени относящиеся к данной проблематике.

Вайнер [5] в 1973 г. предложил специальную конструкцию (дерево префикс-идентификаторов), которая позволяла в линейное время находить самые длинные повторения в тексте. Развив метод Вайнера и использовав новые конструкции (цепи, полисегменты), Слисенко [6] в 1977 г. построил алгоритм для нахождения в "реальное время" всех периодичностей в тексте. Введенная им для оценки сложности вычислений модель адресной машины наиболее точно соответствовала однопроцессорным вычислениям в оперативной памяти. Термин "в реальное время" соответствует тому, что время работы алгоритма после поступления очередной буквы текста, но до поступления следующей не превосходит константы.

Автор совместно с Ю.Г.Косаревым и Т.Н.Титковой в 1975 г. [7] предложил итеративный по l алгоритм отыскания всех повторов, использующий идеи хэш-адресации и специальную конструкцию для хранения информации о предыдущей итерации (информационная лента). Алгоритм ориентирован на работу с текстами, длина которых может п р е в ы ш а т ь размер оперативной памяти S (в битах). При $N < S$ алгоритм имеет по каждой итерации линейную "в среднем" трудоемкость (в том смысле, в котором линейна "в среднем" процедура хэш-адресации [1]). Выявление симметричных участков в тексте сводится к нахождению общих повторов в левой и правой частях конкатенации $T \sqcup T_0$, где T – исходный текст, T_0 – тот же текст, записанный в обратном порядке, $\sqcup$ – символ, отсутствующий в T и T_0.

§4. Меры сходства символьных последовательностей

Возможным обобщением рассмотренной выше методики описания символьных последовательностей является объединение в один таксон не только точно совпадающих 1-грамм, но и 1-грамм, отличающихся друг от друга, однако близких в определенном смысле. В связи с этим представляют интерес возможные варианты определения мер сходства символьных последовательностей. Заметим также, что понятие меры сходства в том или ином виде лежит в основе всех задач классификации, в том числе классификации символьных последовательностей (например, знаков пунктуации в генетических текстах). Рассмотрим три типа мер сходства.

I. Р а н г о в ы е м е р ы с х о д с т в а символьных последовательностей можно строить на основе различных известных определений коэффициента ранговой корреляции. Проиллюстрируем это на примере коэффициента Спирмэна ρ [8].

Пусть u и v - символьные последовательности, $\Phi_1(u)$ и $\Phi_1(v)$ - их частотные характеристики 1-го порядка, в которых 1-граммы расположены по убыванию частоты встречаемости $F_1(r)$, и x_1 - произвольная 1-грамма из алфавита A_1 мощности $R_1 = n^1$. Введем аналог расстояния

$$S_1(u,v) = \sum_{x_1 \in A_1} (r_u(x_1) - r_v(x_1))^2, \qquad (8)$$

где $r_u(x_1)$ и $r_v(x_1)$ - ранги 1-граммы x_1 в обоих упорядочениях. Тогда аналогом коэффициента Спирмэна для характеристик 1-го порядка является значение

$$\rho_1(u,v) = 1 - \frac{6 S_1(u,v)}{R_1(R_1^2 - 1)}, \quad 1 = 1,2,\ldots . \qquad (9)$$

При вычислении (8) группы равночастотных 1-грамм представляются усредненным рангом, а в (9) вносится поправка на "связанность" рангов [8]. Отсутствующие в u или v 1-граммы также образуют равночастотную группу с $F_1(r) = 0$. Если таких 1-грамм много, то естественнее определить параметр R_1 как мощность объединения множеств различных 1-грамм из u и v.

Обобщением (9) на случай m последовательностей ($m \geq 2$) является 1-граммный аналог коэффициента конкордации W [8]:

$$W_l = \frac{12\,V_l}{m^2 R_l (R_l^2 - 1)}, \quad l = 1,2,\ldots, \tag{10}$$

где V_l - сумма квадратов отклонений суммы рангов каждой l-граммы (по всем m упорядочениям) от среднего значения, равного $\frac{1}{2}m(R_l+1)$. При наличии связанных рангов в (10) вводится соответствующая поправка.

2. l-г р а м м н ы е м е р ы с х о д с т в а представляют собой различные модификации известной теоретико-множествен - ной меры сходства в виде отношения пересечения двух множеств к их объединению. В качестве примера приведем меру [9]:

$$\lambda(u,v) = \frac{\sum\limits_{\alpha} \min\{F(u:\alpha), F(v:\alpha)\}\cdot|\alpha|}{\sum\limits_{\alpha} \max\{F(u:\alpha), F(v:\alpha)\}\cdot|\alpha|}, \tag{11}$$

где $F(u:\alpha)$, $F(v:\alpha)$ - частоты встречаемости произвольной l-грам - мы α в u и v соответственно, а $|\alpha| = l$ - длина l-граммы. Эту ме- ру, по-видимому, можно рекомендовать лишь для коротких последова- тельностей, ибо она трудоемка для вычисления и при больших N дает заниженные (с интуитивной точки зрения) значения.

3. М е р ы, о с н о в а н н ы е н а в ы ч и с л е н и и "р е д а к ц и о н н о г о" р а с с т о я н и я, в некотором смысле оперируют уже с "разрывными" l-граммами. Определим расстоя- ние $D(u,v)$ между последовательностями u и v как минимальное чис- ло операций типа "устранение" символа, "добавление" символа или "замена" одного символа другим, требующихся для перевода u в v (или наоборот). В таком виде это расстояние было введено Левенш - тейном, а в [10] было предложено называть его "редакционным" и описывался алгоритм его вычисления.

С расстоянием $D(u,v)$ тесно связано понятие максимально длин- ной общей подпоследовательности двух последовательностей. Будем на- зывать X подпоследовательностью Y, если существует монотонно возрастающая последовательность целых $m_1, m_2, \ldots, m_{N_X}$ $(1 \le m_i \le N_Y)$ такая, что $X[i] = Y[m_i]$ для $1 \le i \le N_X$ (здесь $X[i]$ -i-й эле- мент последовательности X). Далее, X является общей подпоследо - вательностью последовательностей u и v, если X - подпоследова - тельность как u, так и v.

Максимально длинная общая подпоследовательность есть общая подпоследовательность с наибольшим числом членов. Обозначим ее длину через $L(u,v)$. Расстояние $D(u,v)$ связано с $L(u,v)$ соот-

ношением [10]: $D(u,v) = N_u + N_v - 2L(u,v)$. Меру близости между u и v можно определить либо непосредственно через $D(u,v)$:

$$\beta_1(u,v) = 1/(1 + D(u,v)), \qquad (12)$$

либо через $L(u,v)$:

$$\beta_2(u,v) = L(u,v)/\max\{N_u, N_v\}. \qquad (13)$$

Очевидно, что обе меры симметричны, достигают при $u = v$ максимального значения, равного I, и изменяются в пределах:

$$0 < \beta_1(u,v) \leq 1,$$

$$0 \leq \beta_2(u,v) \leq \min\{N_u, N_v\}/\max\{N_u, N_v\}.$$

В отличие от (9) и (II), меры (I2) и (I3) уже учитывают порядок следования элементов. Трудоемкость их вычисления в наихудшем случае составляет $O(N_u \cdot N_v)$.

Выводы

Предложен способ описания символьных последовательностей, ориентированный на выявление периодической структуры текста. В зависимости от решаемой содержательной задачи текст может быть описан с требуемой степенью детализации. Полное описание обеспечивает возможность однозначного восстановления текста. На основе предложенного описания могут быть получены различные интегральные характеристики, играющие важную роль при классификации текстов. Описание инвариантно к перекодировкам алфавита и устойчиво к помехам, описанным в §I.

Отметим, что понятие повтора лежит в основе определения избыточности языка. Роль повторов в объяснении механизмов восприятия человеком звуковых последовательностей неоднократно отмечалась психологами. В терминах повторов могут быть сформулированы задачи выявления симметричных участков в тексте, выявления инверсий и транспозиций в генетических текстах, автоматического обнаружения морфем - элементарных смыслонесущих единиц языка [II].

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. WOODS W.A. Transition network grammars for natural language analysis.- CACM,1970,v.13, N 10,p.591-606.

3. МАНДЕЛЬБРОТ Б. О рекуррентном кодировании, ограничивающем влияние помех. -В кн.: Теория передачи сообщений.М.,1975,с.139-157.

4. ШЕННОН К. Предсказание и энтропия печатного английского текста. -В кн.: Работы по теории информации и кибернетике.М.,1963, с. 669-686.

5. WEINER P. Linear pattern matching algorithms. - In: IEEE 14th Annual Symposium on Switching and Automata Theory, 1973,p.1-11.

6. СЛИСЕНКО А.О. Распознавание предиката вхождения в реальное время. Л., Наука. - 24 с. (Препринт/ЛОМИ:Р-7-77).

7. ГУСЕВ В.Д., КОСАРЕВ Ю.Г., ТИТКОВА Т.Н. О задаче поиска повторяющихся отрезков текста. -В кн.: Вычислительные системы.Вып. 62. (Ассоциативное кодирование.) Новосибирск, 1975, с.49-71.

8. КЕНДЭЛ М. Ранговые корреляции. -М.: Статистика, 1975. - 213 с.

9. FINDLER N.V., LEEUWEN J.van. A family of similarity measures between two strings. - IEEE Trans.on Pattern Analysis and Mach. Intell.,1979, v.PAMI-1, N 1, p.116-118.

10. WAGNER R.A., FISCHER M.J. The string-to-string correction problem.- JACM,1974,v.21, N 1,p.168-173.

11. СУХОТИН Б.В. Оптимизационные методы исследования языка. Автореф. дис. на соиск. учен. степени доктора филол. наук. М., 1979. - 28 с.(Моск.Госуниверситет).

Поступила в ред.-изд. отд.
4 мая 1981 года