АНАЛИЗ ТЕКСТОВ И СИГНАЛОВ
(Вычислительные системы)
1987 год Выпуск 123

УДК 681.3.06

ЧИСЛЕННЫЙ КРИТЕРИЙ БЛИЗОСТИ ТЕКСТОВ

В.А. Леус

В [I] изложен общий подход к заданию критериев близости (расстояний) на множестве произвольных конечных текстов. Он охватывает как частные случаи расстояние Хемминга и обобщенное расстояние Хемминга. Геометрическую структуру на множестве текстов можно вводить различными способами. Пусть расстояние ρ определяет одну геометрическую структуру, а расстояние ρ' - другую. Разрешающая способность расстояния ρ' выше, чем у расстояния ρ, если для любых трех текстов L, M, N неравенство $\rho(L,M) \neq \rho(L,N)$ влечет неравенство $\rho'(L,M) \neq \rho'(L,N)$, но не наоборот. Подход дает возможность достаточно полно учитывать различие и взаиморасположение символов в сравниваемых текстах и тем самым получать высокую разрешающую способность расстояния. Однако комбинаторный характер критериев близости высокого разрешения часто сопряжен с экспоненциальным объемом вычислений.

В настоящей статье предлагается компромиссный критерий близости для текстов. Метод вычисления значения критерия для двух данных текстов использует свойство выпуклости функции, задающей расстояние между отдельными вхождениями букв. Это открывает возможность для направленного поиска, трудоемкость которого оказывается полиномиально зависящей от объемов сравниваемых текстов. Приводятся примеры работы метода в некоторых прикладных задачах.

§I. Расстояние между текстами

Введение геометрической структуры в множестве текстов основывается на понятии контакта двух текстов M и N. Пусть a и b - произвольные буквы алфавита A. Вхождение буквы a на месте μ в текст M обозначим через a_μ, а буквы b на месте ν в текст

$N - b_\nu$. Контакты $K(M,N)$ рассматриваемых текстов M и N есть двудольный граф, одну долю вершин которого образуют вхождения букв в текст M, другую - вхождения букв в текст N, причем каждая вершина обязательно инцидентна одному и только одному ребру. Ребра a_μ и b_ν называются связями, а висячие ребра (a_μ, Λ) (или b_ν, Λ)), где Λ - пустой символ, называются псевдосвязями. Каждой связи и псевдосвязи ставится в соответствие неотрицательное число, называемое ее длиной. При фиксированном способе назначения длин связей и псевдосвязей за длину контакта принимается сумма длин всех его ребер. Расстояние $\rho(M,N)$ между текстами есть минимум длин контактов этих текстов.

Чем больше тонкостей в различии текстов учитывает расстояние ρ, тем обширнее множество допустимых контактов. С другой стороны, чем менее представительно множество контактов, тем легче искать контакт с минимальной длиной (минимальный контакт) для данных текстов. В качестве контактов для текстов M и N будем рассматривать только такие двудольные графы, в которых: а) каждое ребро (связь)

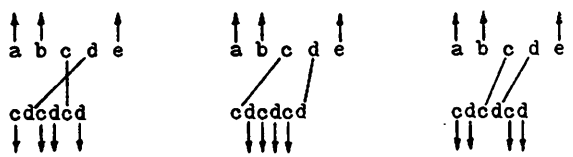

Рис. I

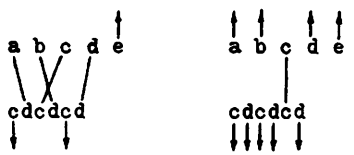

Рис. 2

инцидентно паре (a_μ, a_ν) вхождений одной и той же буквы; б) все непарные вхождения (и только они) являются концами псевдосвязей. На рис. I показаны три разных контакта между текстами. Псевдосвязи обозначены стрелками, обращенными наружу. На рис. 2 даны примеры двудольных графов, которые контактами не являются (в левом нарушено условие "а", в правом - условие "б").

Допустим, что текст M состоит из m вхождений одной буквы, а текст N - из n вхождений той же самой буквы (для определенности $m \le n$). Тогда число μ всевозможных контактов $K(M,N)$ равно $C_n^m \cdot m! = n!/(n-m)!$. Если некая k-я буква алфавита имеет m_k

вхождений в текст M и n_k вхождений в текст N, то можно рассматривать контакты подтекстов (фрагментов текстов) M_k и N_k, состоящих из всех вхождений k-й буквы в тексты M и N. Число таких контактов $K_k(M,N)$ текстов M и N по k-й букве есть $\mu_k = n_k!/|n_k - m_k|!$, откуда количество всех контактов K(M,N) не меньше, чем произведение $\Pi\mu_k$ по всем буквам, имеющим вхождения хотя бы в одно из сравниваемых слов. Мощность множества контактов есть очень быстро растущая функция от числа вхождений букв в сравниваемых текстах, поэтому актуальной оказывается задача сокращения перебора.

Настоящая работа посвящена методу направленного поиска контакта минимальной длины для двух конечных текстов. Основой для построения различных текстов служит целочисленная сетка, в каждом узле которой стоит один экземпляр пустого символа Λ. Текст получается в результате замены пустых символов в конечном числе позиций содержательными буквами алфавита A (все формально строгие определения приведены в [I]). При таком определении два текста, получаемые один из другого сдвигом по сетке, являются разными. Число вхождений содержательных букв в тексте назовем размером (или объемом) текста.

Количества связей и псевдосвязей по любой букве во всех контактах данной пары текстов постоянны. При любом способе назначения длины псевдосвязи сумма длин всех псевдосвязей одинакова для всех контактов данной пары текстов. В длине контакта выделяются две компоненты - суммарная длина связей и суммарная длина псевдосвязей. В этом смысле для обозначения расстояния удобно использовать символику комплексного числа $\rho = \alpha + \beta i$, где часть Re_ρ отвечает вкладу связей, а часть Im_ρ - вкладу псевдосвязей. Расстояние между текстами, имеющими одинаковые количества вхождений каждой буквы, имеет действительное значение, а расстояние любого содержательного текста от пустого - чисто мнимое.

Согласно условию "а" всякий контакт любых двух текстов можно разбить на непересекающиеся "контакты по букве" соответственно числу букв, имеющих вхождения в сравниваемых текстах. В минимальном контакте каждый "контакт по букве" минимален, так что процедура поиска минимального контакта сводится к независимому отысканию минимальных контактов по каждой из букв. Длина минимального контакта по некоторой букве "b" есть р а с с т о я н и е п о б у к в е $\rho_b(M,N)$ между текстами, а полное расстояние получается суммированием величин ρ_b по всем $b \in A$. Таким образом, для вычис -

ления расстояния кардинальной оказывается задача поиска минимального контакта текстов в алфавите A , содержащем одну букву, которую можно обозначить, например, символом 1.

Связь, инцидентную вхождениям 1_μ и 1_ν , будем обозначать (μ,ν), а ее длину - $s_{\mu\nu}$; псевдосвязь и ее длину естественно обозначить (μ,Λ) и s_μ (соответственно (ν,Λ) и s_ν). Величину $s_{\mu\nu}$ будем считать функцией от модуля разности $|\mu - \nu|$ индексов вхождений, а длину псевдосвязей - тождественно равной константе s. Очевидно, что расстояние между текстами обладает свойством симметрии, т.е. $\rho(M,N) = \rho(N,M)$. Если к тому же потребовать, чтобы функция $s_{\mu\nu}(|\mu-\nu|)$ принимала нулевые значения тогда и только тогда, когда $\mu = \nu$, то расстояние будет обладать свойством тождества: $\rho(M,N) = 0 \Leftrightarrow M = N$. Наконец, исключим для длин связей и псевдосвязей отрицательные значения, т.е. действительная и мнимая части расстояния всегда неотрицательны.

Пусть (μ_1,ν_1) и (μ_2,ν_2) - две связи в контакте $K(M,N)$. При $(\mu_2-\mu_1)\cdot(\nu_2-\nu_1)>0$ связи назовем параллельными, в противном случае - скрещивающимися. Для любых двух вхождений μ_1 и μ_2 в тексте M и любых двух вхождений ν_1 и ν_2 в тексте N существуют две пары связей $((\mu_1,\nu_1),(\mu_2,\nu_2))$ и $((\mu_1,\nu_2),(\mu_2,\nu_1))$. Эти пары назовем *сопряженными*. В одной из сопряженных пар связи параллельные, в другой - скрещивающиеся.

ТЕОРЕМА I о парных связях. *Если* $s(x)$ - *монотонная неубывающая выпуклая (вниз) функция, то сумма длин пары параллельных связей не превосходит суммы длин сопряженной пары скрещивающихся связей.*

Для доказательства этой теоремы нам потребуется одно свойство выпуклых функций, которое сформулируем в виде леммы. Пусть $f(\vec{x})$ - выпуклая функция, определенная на R^n, $\vec{x}_1$, и $\vec{x}_2$ - радиус-векторы двух произвольных несовпадающих точек в R^n, t - действительное число $(0 \le t \le 1)$, вектор $\vec{\delta} = t(\vec{x}_2 - \vec{x}_1)$, Δ_1 - приращение функции $f(\vec{x})$ при смещении из точки $\vec{x}_1$ в точку $(\vec{x}_1 + \vec{\delta})$, Δ_2 - приращение функции при смещении из точки $(\vec{x}_2 - \vec{\delta})$ в точку $\vec{x}_2$.

ЛЕММА. *Для выпуклой вниз функции* $f(\vec{x})$ *справедливо неравенство* $\Delta_1 = f(\vec{x}_1+\vec{\delta}) - f(\vec{x}_1) \le f(\vec{x}_2) - f(\vec{x}_2-\vec{\delta}) = \Delta_2$, *а для выпук-*

лой вверх-обратное неравенство $\Delta_2 \le \Delta_1$.

ДОКАЗАТЕЛЬСТВО леммы. Значение функции $f(\vec{x})$ в точке $\vec{x} = (\vec{x}_1 + \vec{\delta})$ есть

$$f(\vec{x}_1 + \vec{\delta}) = f(\vec{x}_1 + t\vec{x}_2 - t\vec{x}_1) = f((1-t)\vec{x}_1 + t\vec{x}_2) .$$

Согласно определению выпуклости вниз имеем неравенство

$$f((1-t)\vec{x}_1 + t\vec{x}_2) \le (1-t)f(\vec{x}_1) + tf(\vec{x}_2) .$$

Значение функции $f(\vec{x})$ в точке $\vec{x} = (\vec{x}_2 - \vec{\delta})$ по определению выпуклости вниз оценивается следующим образом:

$$f(\vec{x}_2 - \vec{\delta}) = f(\vec{x}_2 - t\vec{x}_2 + t\vec{x}_1) = f(t\vec{x}_1 + (1-t)\vec{x}_2) \le tf(\vec{x}_1) + (1-t)f(\vec{x}_2) .$$

Отсюда выводим оценку для суммы значений

$$f(\vec{x}_1 + \vec{\delta}) + f(\vec{x}_2 - \vec{\delta}) \le (1-t)f(\vec{x}_1) + tf(\vec{x}_1) + tf(\vec{x}_2) + (1-t)f(\vec{x}_2) = f(\vec{x}_1) + f(\vec{x}_2) .$$

Или, перенося слагаемые, получим

$$\Delta_1 = f(\vec{x}_1 + \vec{\delta}) - f(\vec{x}_1) \le f(\vec{x}_2) - f(\vec{x}_2 - \vec{\delta}) = \Delta_2 .$$

Обратное неравенство $\Delta_2 \le \Delta_1$ для функции $f(\vec{x})$, выпуклой вверх, получается аналогичным образом. Лемма доказана, переходим к доказательству теоремы о парных связях.

Пусть для определенности $\mu_1 < \mu_2$, $\nu_1 < \nu_2$, так что (μ_1, ν_1) и (μ_2, ν_2) - параллельные, а (μ_1, ν_2) и (μ_2, ν_1) - скрещивающиеся связи. Введем обозначения: $c = |\mu_1 - \nu_1|$, $d = |\mu_2 - \nu_2|$, $e = |\mu_1 - \nu_2|$, $f = |\mu_2 - \nu_1|$, $P = e + f - c - d$. Найдем значения величины P при различных взаиморасположениях индексов μ_1, μ_2, ν_1, ν_2 на числовой оси:

1. $\nu_1 < \nu_2 \le \mu_1 < \mu_2$:

$$P = \mu_1 - \nu_2 + \mu_2 - \nu_1 - \mu_1 + \nu_1 - \mu_2 + \nu_2 = 0;$$

2. $\nu_1 \le \mu_1 \le \nu_2 \le \mu_2$:

$$P = \nu_2 - \mu_1 + \mu_2 - \nu_1 - \mu_1 + \nu_1 - \mu_2 + \nu_2 = 2(\nu_2 - \mu_1) \ge 0;$$

3. $\mu_1 \le \nu_1 < \nu_2 \le \mu_2$:

$$P = \nu_2 - \mu_1 + \mu_2 - \nu_1 - \nu_1 + \mu_1 - \mu_2 + \nu_2 = 2(\nu_2 - \nu_1) > 0;$$

4. $\nu_1 \le \mu_1 < \mu_2 \le \nu_2$:

$$P = \nu_2 - \mu_1 + \mu_2 - \nu_1 - \mu_1 + \nu_1 - \nu_2 + \mu_2 = 2(\mu_2 - \mu_1) > 0;$$

5. $\mu_1 \le \nu_1 \le \mu_2 \le \nu_2$:

$$P = \nu_2 - \mu_1 + \mu_2 - \nu_1 - \nu_1 + \mu_1 - \nu_2 + \mu_2 = 2(\mu_2 - \nu_1) \ge 0;$$

6. $\mu_1 < \mu_2 \le \nu_1 < \nu_2$:

$$P = \nu_2 - \mu_1 + \nu_1 - \mu_2 - \nu_1 + \mu_1 - \nu_2 + \mu_2 = 0.$$

Все возможности исчерпаны, и всякий раз $P \ge 0$, т.е. всегда

$$c + d \le e + f. \qquad (1)$$

Дальнейшие рассуждения будем проводить в предположении, что $c \le d$. Обратное неравенство рассматривать нет необходимости в силу симметрии длин параллельных и скрещивающихся связей относительно изменения направления возрастания индексов на противоположное.

При $\nu_2 \le \mu_2$ $d = \mu_2 - \nu_2$ и заведомо $\nu_1 < \mu_2$, так что $f = \mu_2 - \nu_1$. Имеем $f - d = \mu_2 - \nu_1 - \mu_2 + \nu_2 = \nu_2 - \nu_1 > 0$, т.е.

$$d < f. \qquad (2)$$

При $\mu_2 \le \nu_2$ $d = \nu_2 - \mu_2$ и заведомо $\mu_1 < \nu_2$ и $e = \nu_2 - \mu_1$. Поэтому $e - d = \nu_2 - \mu_1 - \nu_2 + \mu_2 = \mu_2 - \mu_1 > 0$, откуда

$$d < e. \qquad (3)$$

Пусть $f \le e$, а $\nu_2 \le \mu_2$. Тогда $f = \mu_2 - \nu_1$, $d = \mu_2 - \nu_2$. Если допустить, что $\mu_1 \le \nu_1$, то $e = \nu_2 - \mu_1$, $c = \nu_1 - \mu_1$ и $(e - f) = (\nu_2 - \mu_1) - (\mu_2 + \nu_1) = (\nu_1 - \mu_1) - (\mu_2 - \nu_2) = (c - d) \le 0$. Таким образом, здесь совместимым с условием $f \le e$ является только равенство $f = e$. Но тогда согласно (2) имеем

$$d < f = e. \qquad (4)$$

Предположим, что $\nu_1 < \mu_1$. Тогда $c = \mu_1 - \nu_1$ и либо $\mu_1 \le \nu_2$, либо $\nu_2 < \mu_1$.

При $\mu_1 \le \nu_2$ имеем $e = \nu_2 - \mu_1$, и

$$(e-f) = (\nu_2 - \mu_1) - (\mu_2 - \nu_1) = -(\mu_1 - \nu_1) - (\mu_2 - \nu_2) = -(c+d) < 0.$$

При $\nu_2 < \mu_1$ имеем $e = \mu_1 - \nu_2$ и получаем неравенство

$$(e-f) = (\mu_1 - \nu_2) - (\mu_2 - \nu_1) = (\mu_1 - \mu_2) + (\nu_1 - \nu_2) < 0.$$

В обоих случаях оказывается $e < f$, значит, неравенство $\nu_1 < \mu_1$ несовместимо с условием $f \le e$. Таким образом, согласно (3) и (4), получается, что условия $c \le d$ и $f \le c$ влекут за собой неравенство $d < e$.

Пусть теперь $e \le f$ и $\mu_2 < \nu_2$. Тогда $e = \nu_2 - \mu_1$ и $d = \nu_2 - \mu_2$. Если предположить, что $\nu_1 \le \mu_1$, то $f = \mu_2 - \nu_2$, $c = \mu_1 - \nu_1$, откуда

$$(f-e) = (\mu_2 - \nu_1) - (\nu_2 - \mu_1) = (\mu_1 - \nu_1) - (\nu_2 - \mu_2) = (c-d) \le 0,$$

Здесь с условием $e \le f$ совместимо только равенство $e = f$, и согласно (3) имеем неравенство

$$d < e = f. \qquad (5)$$

Предположим, что $\mu_1 < \nu_1$, тогда $c = \nu_1 - \mu_1$ и либо $\nu_1 \le \mu_2$, либо $\mu_2 < \nu_1$. При $\nu_1 \le \mu_2$ имеем $f = \mu_2 - \nu_1$, откуда

$$(f-e) = \mu_2 - \nu_1 - \nu_2 + \mu_1 = (\mu_1 - \nu_1) + (\mu_2 - \nu_2) = -c - d < 0.$$

При $\mu_2 < \nu_1$ имеем $f = \nu_1 - \mu_2$, откуда

$$(f-e) = \nu_1 - \mu_2 - \nu_2 + \mu_1 = (\mu_1 - \mu_2) + (\nu_1 - \nu_2) < 0 .$$

Наше последнее предположение приводит, таким образом, к противоречию с условием $e \le f$. Остается либо принять неравенство $\nu_1 \le \mu_1$, из которого следует (5), либо отказаться от неравенства $\mu_2 < \nu_2$, что приводит к (3). Из (3) и (5) вытекает, что условия $c \le d$ и $e \le f$ влекут за собой неравенство $d < f$.

Общий итог проведенных рассуждений таков. Как бы ни были расположены на числовой оси индексы четырех вхождений в сопряженных парах скрещивающихся и параллельных связей, модуль разности концевых вхождений максимален у скрещивающейся связи. Принимая во внимание это обстоятельство, нетрудно получить утверждение теоремы.

Рассмотрим случай, когда $e \le f$, и, по доказанному, $d < f$. Здесь имеются три возможности упорядочения величин c, d, e и f.

Пусть $e \le c \le d \le f$, тогда из (I) имеем

$$0 \le c - e \le f - d = p. \qquad (6)$$

Поскольку функция $s(x)$ по условию теоремы монотонная неубывающая, то, учитывая (6), получим

$$s(c) - s(e) = s(c - e + e) - s(e) \le s(p + e) - s(e).$$

В силу выпуклости вниз функции $s(x)$ имеем по доказанной лемме

$$s(e + p) - s(e) \le s(f) - s(f - p) = s(f) - s(f - f + d) = s(f) - s(d).$$

Поэтому $s(c) - s(e) \le s(f) - s(d)$ или, что то же, $s(c) + s(d) \le s(e) + s(f)$.

Пусть теперь $c \le e \le d \le f$. Дважды используя только монотонное неубывание функции $s(x)$, получаем

$$s(c) + s(d) \le s(e) + s(d) \le s(e) + s(f).$$

Пусть, наконец, $c \le d \le e \le f$. Поскольку c и d по отдельности и вместе не превосходят величин e и f, то в силу монотонного неубывания $s(x)$ очевидно, что

$$s(c) + s(d) \le s(e) + s(f).$$

Случай, когда $f \le e$ и $d < e$, также допускает три возможности, которые мы кратко рассмотрим по аналогии.

а) $f \le c \le d < e$. Из (I) имеем

$$0 \le c - f \le e - d = q. \qquad (7)$$

Используя неубывание $s(x)$ и учитывая (7), имеем

$$s(c) - s(f) = s(c - f + f) - s(f) \le s(q + f) - s(f),$$

а в силу выпуклости вниз получим

$$s(f + q) - s(f) \le s(e) - s(e - q) = s(e) - s(d).$$

Отсюда $s(c) - s(f) \le s(e) - s(d)$ или, что то же, $s(c) + s(d) \le s(e) + s(f)$.

б) $c \le f \le d < e$. В силу неубывания $s(x)$ имеем

$$s(c) + s(d) \le s(f) + s(d) \le s(f) + s(e).$$

в) $c \le d \le f \le e$. Также в силу неубывания

$$s(c) + s(d) \le s(f) + s(e).$$

Во всех возможных случаях мы убедились в том, что по сумме длин параллельные связи не превосходят скрещивающихся связей сопряженной пары. Теорема доказана.

§2. Алгоритм вычисления расстояния

Рассмотрим тексты M и N, состоящие из вхождений одной и той же буквы. Для определенности примем, что число вхождений в текст M не превосходит числа вхождений в текст N, т.е. $m \le n$. Пусть M' есть некоторый подтекст текста $M (m' < m)$. Имеет смысл говорить о расстоянии $\rho(M', N)$ между M', рассматриваемым как отдельный текст, и текстом N. Пусть контакт K' реализует это расстояние, а π' - множество тех концов его связей, которые входят в N. Назовем π' множеством, ближайшим к подтексту M' (для краткости - просто ближайшим).

Всякий текст является объединением (в теоретико-множественном смысле) своих подтекстов. Выберем непересекающиеся подтексты $M_1, M_2, \ldots, M_I$ так, что $M' = \bigcup_{i=1}^{I} M_i$. Обозначим через $\pi_1, \pi_2, \ldots, \pi_I$ множества, ближайшие к $M_1, M_2, \ldots, M_I$ соответственно. Эти ближайшие могут иметь непустые пересечения между собой. В [I] доказано существование такого π', ближайшего к подтексту M', которое включает в себя объединение ближайших, т.е. $\pi' \supseteq \bigcup_{i=1}^{I} \pi_i$.

Пусть N' – подтекст текста N, причем γ – наименьший, δ – наибольший индексы вхождений этого подтекста. Подтекст N' называется сегментом текста N, если всякое вхождение a_ν текста N при $\gamma \le \nu \le \delta$ является также и вхождением подтекста N'. В [I] показано, что если объединение ближайших $\bigcup_i \pi_i$ есть сегмент, то среди содержащих его ближайших к объединению $\bigcup_i M_i$ найдется сегмент. Эти два утверждения из [I] и сформулированная в предыдущем параграфе теорема о парных связях положены в основу метода направленного поиска минимального контакта двух текстов.

Алгоритм поиска минимального контакта сводится к повторяющемуся процессу отыскания ближайшего к объединению подтекстов M_i. На первом этапе работы алгоритма за исходное принимается разбиение текста M на отдельные вхождения. Для каждого подтекста M_i (состоящего здесь из одного вхождения a_μ) ищется ближайшее из вхождений a_ν текста N. В силу условия теоремы о парных связях длина связи $s_{\mu\nu}$ как функция $s_\mu(\nu)$ переменного индекса ν при фиксированном индексе μ является выпуклой вниз функцией, поэтому вычисленные для двух соседних аргументов значения $s_\mu(\nu_1)$ и $s_\mu(\nu_2)$ этой функции указывают направление ее убывания. Минимальная связь выявляется сменой убывания функции $s_\mu(\nu)$ на возрастание. Когда найдены все ближайшие π_i (на первом этапе состоящие из одного вхождения каждое), может оказаться, что они попарно не пересекаются, тогда искомый минимальный контакт состоит из найденных минимальных связей. В противном случае объединение всех π_i разбивается на минимально возможное число J сегментов seg_j ($j = 1,2,\ldots,J<I$). Вхождения сегмента seg_j связаны с вхождениями текста M, образующими некоторый подтекст M_j. Подтексты M_j попарно не пересекаются, исчерпывают все M, и их совокупность берется в качестве исходного разбиения ($\bigcup_j M_j = M$) для второго этапа работы алгоритма.

Второй этап состоит в повторении циклов отыскания ближайшего к объединению при известном объединении ближайших. Любой подтекст M_j, в свою очередь, состоит из подтекстов M_{ij}, для каждого из которых в предыдущем цикле алгоритма было найдено ближайшее π_i (используем опять индексацию по i, хотя это уже другие подтексты). Задача состоит в отыскании ближайшего π_j к объединению $\bigcup_i M_{ij} = M_j$. Поскольку ближайшее к объединению содержит объединение ближайших, то минимальный контакт подтекста M_j с текстом N нет необходимости искать среди всех возможных контактов. Область поиска можно еще сузить, если учесть, что каждое $\bigcup_i \pi_i$ есть сегмент seg_i, и среди содержащих его ближайших к объединению найдется сегмент, т.е. достаточно просматривать только такие контакты, вершины которых в тексте N образуют сегменты. Наконец, в силу теоремы о парных связях минимальный контакт следует искать среди параллельных контактов, содержащих только параллельные связи. Если подтекст $M_j = \bigcup_i M_{ij}$ содержит m_j вхождений, а объединение ближайших $\bigcup_j \pi_j$ содержит n_j вхождений ($n_j \le m_j$), искомый минимальный контакт находится среди $(m_j - n_j) + 1$ параллельных контактов по числу сегментов с m_j вхождениями, содержащими сегмент из n_j вхождений. Множество этих контактов обозначим K_j. На рис. 3 для случая $m_j = 7$, $n_j = 3$ показаны два "зацепленных" сегмента из пяти возможных.

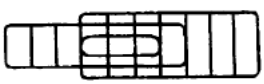

Рис. 3

В множестве K_j возможен направленный поиск контакта минимальной длины. Все контакты в K_j имеют одни и те же вершины в тексте M, и каждый из них индивидуализирован младшим индексом $\tilde{\nu}$ своей вершины из числа вхождений текста N. Все множество K_j упорядочивается по возрастанию $\tilde{\nu}$. Длина контакта $K(\tilde{\nu})$ складывается из длин его параллельных связей. Длина каждой связи по условию теоремы I является выпуклой вниз функцией от ν, и так как эти связи входят в параллельные контакты, вершины которых в тексте M фиксированы, длина каждой связи есть выпуклая вниз функция и от $\tilde{\nu}$. Это обусловливает возможность направленного перебора множества K_j: вычисление длин контактов $K(\nu_1)$ и $K(\nu_2)$ для двух соседних значений младших индексов указывает направление убывания длины. Смена убывания на возрастание говорит о том, что минимальный контакт найден. Возможно применение и более эффективных приемов поиска экстремума выпуклой функции.

После того как в каждом K_j найден минимальный контакт, совокупность всех их вхождений в текст N образует объединение ближайших к подтекстам M_j. Если число этих вхождений равно m, т.е. найденные ближайшие попарно не пересекаются, то объединение минимальных контактов по всем j и дает искомый минимальный контакт для текстов M и N.

На рис.4 приведен пример работы алгоритма при поиске минимального контакта текстов M(m=39) и N(n=40). Полностью связи показаны лишь для первого этапа, а в циклах второго этапа, чтобы не загромождать рисунок, связи показаны частично. Минимальный контакт длиной 255 найден на шестом цикле второго этапа. Имеется в виду действительная часть расстояния, а мнимая часть здесь равняется единице (40 -39 = I).

Трудоемкость поиска объединения ближайших на первом этапе работы алгоритма пропорциональна произведению $m \times n$, так как для каждого вхождения в тексте M достаточно перебрать все вхождения в текст N, чтобы отыскать ближайшее. Для одноциклового второго этапа в худшем случае объединение ближайших для всех вхождений в текст M состоит из одного вхождения в текст N. Тогда множество $K_j(j=1)$ имеет мощность $(m-1)+1=m$ и минимальный контакт отыскивается перебором длин m контактов по m связей в каждом.

В многоцикловом варианте второго этапа подтексты $M_1, M_2, \ldots, M_J$ имеют соответственно количества вхождений $m_1, m_2, \ldots, m_J$. В худшем случае на каждом цикле второго этапа происходит присоединение только одного из J исходных ближайших. Сначала оказываются слитыми ближайшие для j = I и j = 2, и минимальный контакт в первом цикле ищется самое большее либо среди $(m_1+m_2)-m_1+1$ при $m_2 < m_1$, либо среди $(m_1+m_2)-m_2+1$ при $m_1 < m_2$ контактов с (m_1+m_2) связями каждый. Трудоемкость здесь не превосходит величины $(m_1+m_2)\times[m_2+1]$. Во втором цикле к объединению $\pi_1 \cup \pi_2$ присоединяется π_3 и минимальный контакт ищется либо среди $(m_1+m_2+m_3)-(m_1+m_2)+1$ контактов, если $m_3 < (m_1+m_2)$, либо среди $(m_1+m_2+m_3)-m_3+1$ контактов, если $(m_1+m_2) < m_3$, с числом связей $(m_1+m_2+m_3)$. Во всех случаях трудоемкость не превосходит величину, пропорциональную $(m_1+m_2+m_3)\cdot[m_3+1]$. По аналогии трудоемкость второго этапа можно оценить величиной

$$(m_1+m_2)\cdot[m_2+1] + (m_1+m_2+m_3)\cdot[m_3+1] + \ldots + \sum_1^J m_j[m_J+1].$$

Мажорируя каждое слагаемое в круглых скобках суммой всех m_j и вы-

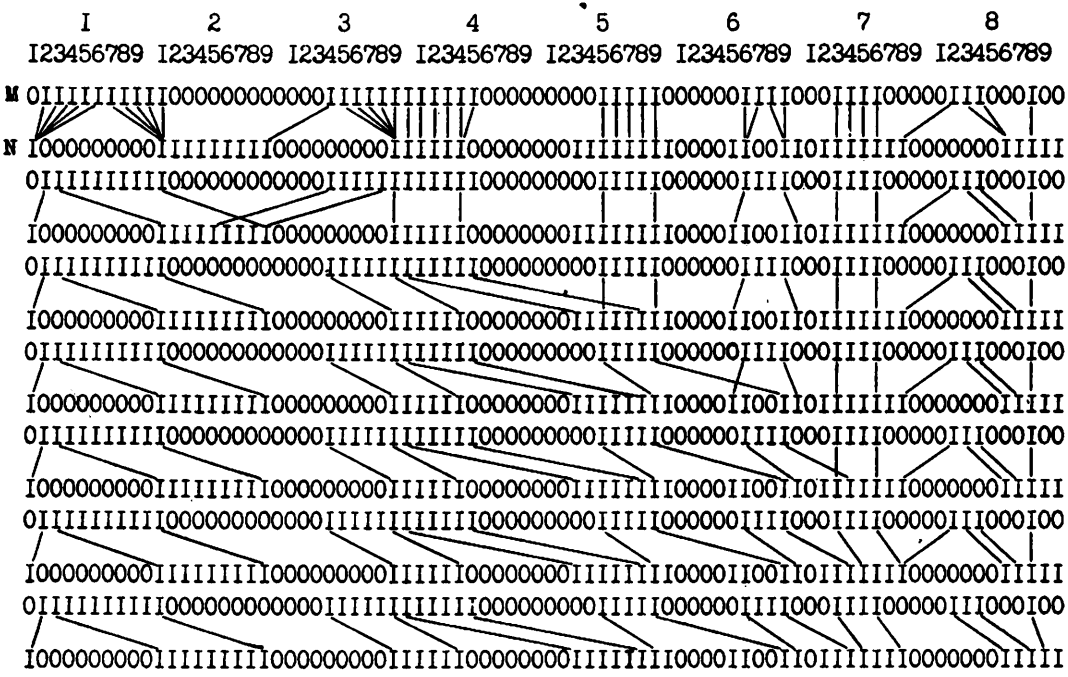

Рис. 4. Пример работы алгоритма при поиске минимального контакта текстов.

нося ее, получим выражение $[m_2 + m_3 + \ldots + m_J + m - 1] \cdot \Sigma m_j$, которое очевидным образом оценивается сверху величиной $2\Sigma m_j \cdot \Sigma m_j = 2m^2$. Таким образом, доказана

ТЕОРЕМА 2 о трудоемкости вычисления расстояния между текстами. Алгоритм вычисления расстояния для двух текстов с m и n вхождениями имеет по числу операций не более чем билинейную по m и n трудоемкость.

§3. Практическое применение

Рассмотрим несколько примеров сравнения текстов, описывающих структуры различной природы. Длину псевдосвязи считаем единичной, а за длину связи принимаем величину модуля разности индексов связанных вхождений.

Если индексы всех содержательных вхождений текста M_0 изменить на одну и ту же величину Δ (смещение), то получится новый текст M_Δ, который мы будем рассматривать как результат сдвига текста M_0 на величину смещения (влево - отрицательное, вправо - положительное). Для двух сравниваемых текстов M_0 и N_0 расстояние $\rho(M_\Delta, N_0)$ равно расстоянию $\rho(M_0, N_{-\Delta})$. Условимся "двигать" текст с меньшим числом вхождений, т.е. считать $N(m \le n)$ неподвижным.

ПРИМЕР I. Для текстов $M(m = 39)$ и $N(n = 40)$, показанных на рис.4, были вычислены расстояния при различных смещениях слова M в пределах от -3 до I7 с единичным шагом. На рис.5 слева приведен график изменения действительной части в зависимости от смещения, откладываемого по оси абсцисс. При нулевом смещении действительная часть принимает значение 255, для получения которого и был реализован процесс поиска минимального контакта, изображенный на рис.4. Приведенный график содержит характерную особенность зависимостей $Re_\rho(\Delta)$, построенных для любой пары текстов, а именно неограниченный рост при стремлении к бесконечности абсолютной величины смещения. Достаточно сдвинуть текст M вправо так, чтобы его наименьший индекс превысил наибольший индекс текста N, как все связи минимального контакта при дальнейших смещениях будут только удлиняться. Аналогичная ситуация достигается и при сдвигах влево. Где-то в промежуточном положении имеется расстояние с минимальной действительной частью, называемое расстоянием между свободными

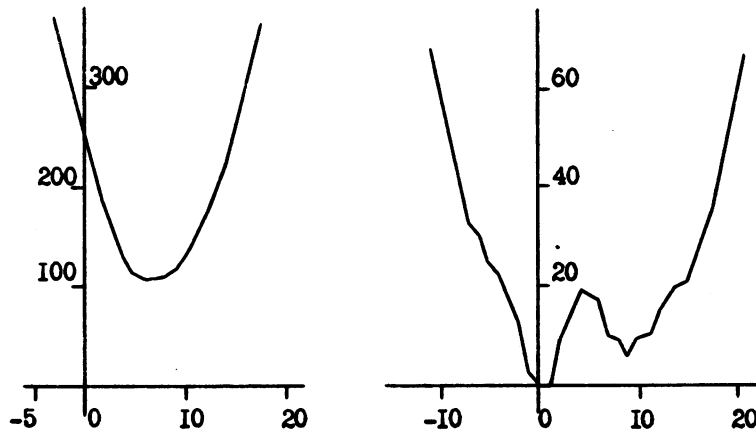
300
200
100
-5
0
10
20
60
40
20
-10
0
10
20

Рис. 5

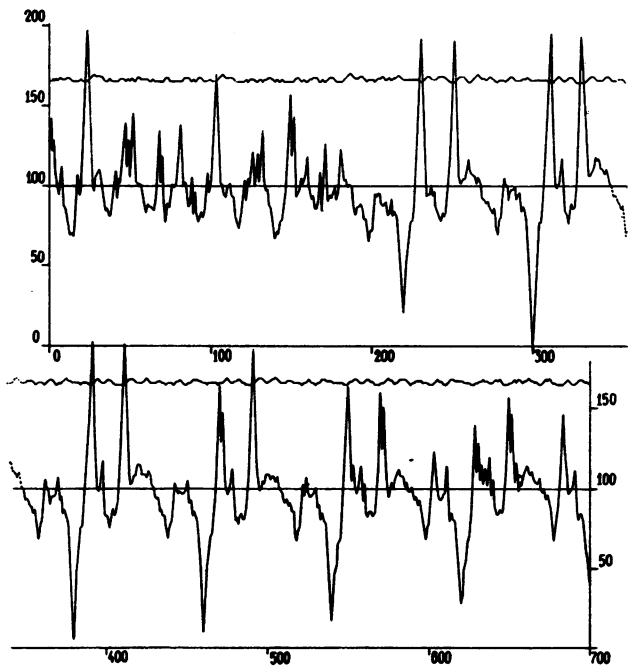
200
150
100
50
0
0
100
200
300
150
100
50
400
500
600
700

Рис. 6

текстами M и N. В нашем примере это расстояние со значением $Re_\rho \leq 105$ достигается при смещении $\Delta = 6$.

ПРИМЕР 2. Зависимость $Re_\rho(\Delta)$ необязательно имеет вид выпуклой (вниз) кривой с единственным минусом. Так, для текстов M(III00000III00000III) и N(IIII000IIIII000IIIII0000000IIII) на рис.5 справа приведена зависимость $Re_\rho(\Delta)$, которая имеет два минимума, разделенных максимумом, равным 19, в точке $\Delta = 4$. В точке $\Delta = 9$ находится локальный минимум со значением 6. Глобальный минимум с нулевым значением достигается в двух соседних точках $\Delta = 0$ и $\Delta = 1$, а мнимая часть $Im_\rho = 9$. Расстояние между текстами в этом примере имеет, таким образом, чисто мнимое значение. Это является формальным выражением того факта, что текст $M(\Delta)$ при определенном смещении оказывается подтекстом текста M.

ПРИМЕР 3. В табл.1 приведены короткий текст M и длинный текст N, вдоль которого перемещается короткий (вверху для удобства показана нумерация позиций). При каждом смещении из текста $M(\Delta)$ "вырезается" связный кусок - окно, включающее все позиции текста $M(\Delta)$. Вхождения текста N, попадающие в окно, образуют слово $\tilde{N}$. Размеры окна могут быть расширены за пределы слова $M(\Delta)$ так, чтобы выполнялось неравенство $m < n$. Для различных смещений вычислялось расстояние между текстами N и $M(\Delta)$. Результаты графически представлены на рис.6. Кривая большой амплитуды - график действительной части, а вверх вынесен график мнимой части, которая (при выбранном окне) изменяется в пределах от 0 до 5. Начиная со смещения $\Delta = 200$, на кривой $Re_\rho(\Delta)$ прослеживается повторяемость с периодом, равным 80.

ПРИМЕР 4. Осадконакопление в морском бассейне является существенно неравномерным процессом. Сносимый с разрушающихся гор в море обломочный материал накапливается сначала в верхней части материкового склона, а когда масса превысит некоторую критическую величину, срывающийся мутьевой поток переотлагает материал на дно бассейна. История осадконакопления фиксируется в виде последовательности чередующихся слоев различных по толщине и зернистости. Для кодирования структуры Большекаройской свиты - древнего осадочного района [2] - была выбрана линейная единица, равная толщине минимального слоя, а все слои были разбиты по составу на семь типов, обозначаемых буквами. Структура разреза описывается текстом, в котором каждому слою поставлено в соответствие столько вхождений буквы данного типа, сколько линейных единиц укладывается на толщине слоя.

Таблица I

	123456789	123456789	123456789	123456789	123456789	123456789	123456789	123456789
M	00I000I000000IIII00I0III0000000001IIII0000IIIII000I0II00000I0000III0000I00IIII00							
N	I0000I000IIII000000I00000III0IIII0000000000000IIIIIII0I000I00000I0I00II000II0I0I							
	0I000I0000IIII00000I0000III000IIII0000000000IIIII0III00000I0000II000I0I00III00I							
	0I0000I0000IIII0000I000III00000IIII0000000000IIIII00III00000I0000II00I00I00III0I0							
	00I000I00000IIII000I00III0000000IIIII000000IIIII000III00000I0000II0I000I00IIII00							
	00I000I000000IIII00I0III0000000000IIIII0000IIIII000I0II00000I0000III0000I00IIII00							
	0I0000I000000IIII000IIII0000000000IIIII0000I0III000I0I0I0000I0000III000000I0IIII00							
	I00000I000000IIII000III0I000000000IIIII0000I0IIII000I00II0000I0000II0I000000I0III0I0							
	I0000I000000I0III000II0II000000000IIIII0000II0III000I00II0000I0000II0I000000I0II00I							
	I000I0000000I0II0I000I0III0000000I0III0I00III0I0000I0000III00000IIII00I							
	I00I0000000I0I0I0I00I0II0I000000I0II0I1001IIII0I00000III0I0000000001II0000III00II							

В табл.2 приведен один из участков, где присутствуют слои типов A,B,C. К четвертому типу были отнесены слои неопределенной природы, которые кодировались нулем, соответствующим пустому символу. Этот участок рассматривался в качестве текста M, а в качестве текста N взят некоторый фрагмент из M длиной в 80 позиций, который смещался вдоль M. Для каждого положения текста N восьмидесятипозиционным окном из M "вырезался" фрагмент, который и сравнивался с текстом N. На рис.7 изображены зависимости (от смещения Δ) общего расстояния $\rho(N,M')$ (жирная кривая) и расстояний по буквам (пунктир - A, тонкая сплошная - B, пунктир с точкой - C). При смещении $\Delta = 600$ все они проходят через ноль, так как фрагмент N взят из текста M именно здесь.

На рис.8 приведены результаты локального сравнения двух текстов. В качестве одного из них взят текст M, а второй получен периодическим повторением того же, что и в предыдущем случае, фрагмента N. Окно на 80 позиций смещалось вдоль этих текстов, и для вырезаемой пары слов вычислялось расстояние. Зависимости расстояний от смещения даны на рис.8 в той же символике, что на рис. ?. Естественно, соответствующие кривые совпадают через каждые 80 позиций.

ПРИМЕР 5. Последовательностями символов описываются биомолекулы. С точки зрения вычисления критерия близости была рассмотрена транспортная рибонуклеиновая кислота (тРНК), молекулярная структура которой приведена в [3]. Эта переносящая аминокислоту валин тРНК имеет сравнительно небольшую молекулу из 77 нуклеотидных звеньев, которые кодируются буквами А,Г,Ц,У. В молекулу входят также редко встречающиеся так называемые неканонические нуклеотиды, которым сопоставлены символы нуля. Закодированная таким образом структура тРНК представлена в табл.3.Для сравнения биологических структур имеет значение не только близость по схожести, но и близость по комплементарности, когда нулевое расстояние приписывается парам разных символов. В нашем примере взаимно комплементарными являются А-У и Г-Ц.

В качестве текста N взята закодированная (табл.3) структура валиновой тРНК. Начальный участок этой структуры, включающий семь звеньев,имеет вид ГАУУУЦГ, а комплементарный ему код ЦУАААГЦ взят в качестве текста M. На рис. 9 точками показана зависимость действительной части расстояния $\rho(M,N)$ от смещения текста M. В

Таблица 2

I23456789 I23456789 I23456789 I23456789 I23456789 I23456789 I23456789 I23456789

N AAAAAAAAAAABBBBBBBBBBBBAAAAAAAAAAAAAACCCCCCCCCCAAAAACCCCCCCCAAAAAAAACCCCCCCCCAAABBBBA

M AAAAAOAAAAAOOOOAAAOAAAAAAAAAAOOAAAAAAAAABBBBBAAAABBBBBAAAAAABBBBBBBBBBAAAAAAAOOAA
AAAAOOOOAAAAABBBBBAAAAAAAABBBBBAAACCCCCCCCCCCAAAAACCCCCCCCCCOOOOOCCCCCCCCCAAAAACCCC
CCCCAAAAAAAOOOAAAAAAAACCCCCCCCCCOOOOOOOOOOAAAAAAAAAAAACCCCCCCCCCCCCOOOOOOOOCCCCCCCC
CCCCOOOOOOBBBBBBAAAAAAAAABCCCCCCCCCCCOOOOOAAAAAAAAAAAAAAAAAAAACCCCCCCCAAAAAAAAAAA
AOOAAAAAAAAAABBBBBBBBAAAAAAAAABBBBBBBAAAAAAAAAAAAOOOOOAAAAAAAAAAAAAACCCCCCCCCAAAAA
AAACCCCCCCCCCCCCOOOOOOOCCCCCCCCCAAAAAAAACCCCCCCCCCOOOOOOOOCCCCCCCCAAAAAAAAAAAAAC
CCCCCCAAAAAAAAAACCCCCCAAAAAAACCCCCCCCOOOOOOOOOCCCCCCCCCAAAAAAAAAAAAAACCCCCCCCCA
AAAAAAACCCCCCCCCCCCAAAAAAAACCCCCCCCCAAAAAAAAAAAAAAABBBBBBBBBBBBAAAAAAAAAAAACCCCCC
CCCAAAAACCCCCCCCAAAAAAAACCCCCCCCAAABBBBAAAABBBBBAAAAAABBBBBBAAAAAAAAAAABBBBBBBBBB
AAAAAAAABBBBBAAAAAAAAAABBBBBBAAAAAAAAABBBBBBAAAAAAAAAACCCCCCCCAAAAAAAAAACCCCCCCC
CCOOOOBBBBBBBBBBBBOOOOCCCCCCCCAAAAAAAACCCCCCCCCOOOOCCCCCCCCCAAAAAAAACCCCCCCCCOO
CCCCCCCCCOOOOOOOOOOOOAAAAAAAAAACCCCCCCCCAAAAAAAABBBBBBBBBBAAAAAAAAACCCCCCCCCAAA
AAAAAACCCCCCCCCAAAAAACCCCCCCCAAAAAAAABBBBBBBAAAAAAAAAAACCCCCCCCCAAAAAAAACCCCCCCC
CCCAAAAAAAABBBBBBBBBBAAAAAAAAABBBBBAAAAAAAAAABBBBBBBBBBBAAAAAAABBBBBBBAAAAAAAACC

Таблица 3

I23456789 I23456789 I23456789 I23456789 I23456789 I23456789 I23456789

ГАУУУЦГУОГУЦОАГОЦГГООАУГАЦАОЦУГЦОУОАЦАЦГЦАГААЦУОЦЦЦАГГОЦГОЦЦЦУГГГГЦГААЦЦАЦЦА

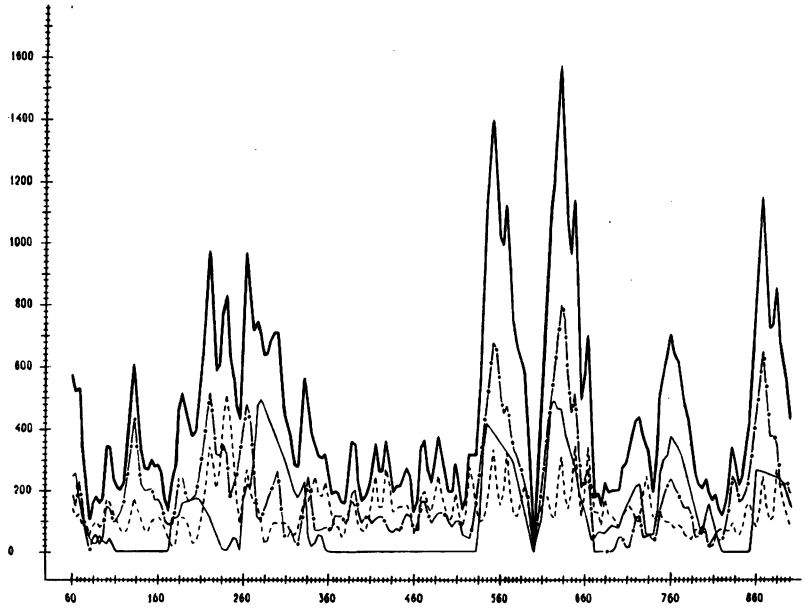

Рис. 7. Зависимость общего расстояния и расстояния по буквам от смещения Δ.

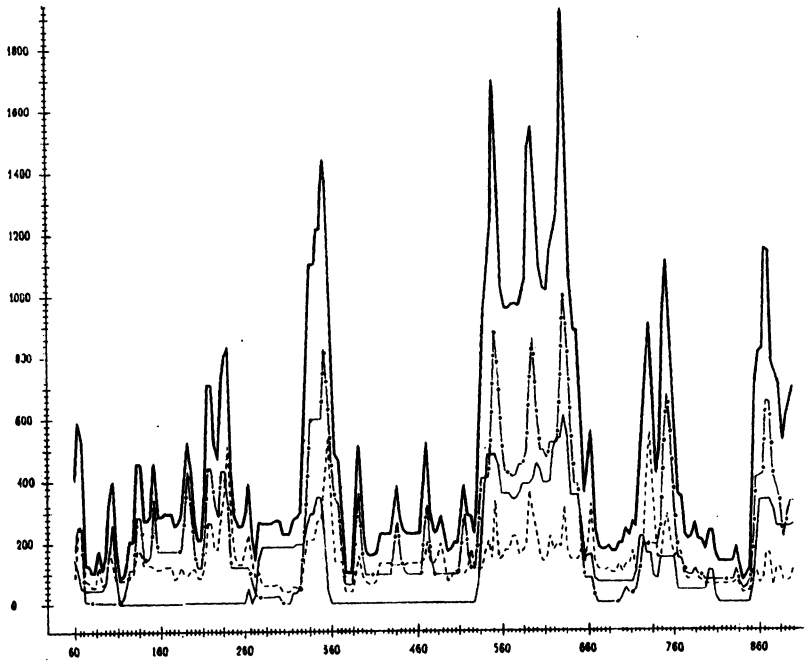

Рис.8. Локальные сравнения двух текстов.

положениях $\Delta = 34$ и $\Delta = 66$ на этой кривой имеются минимумы, но нулевое значение в них не достигается.

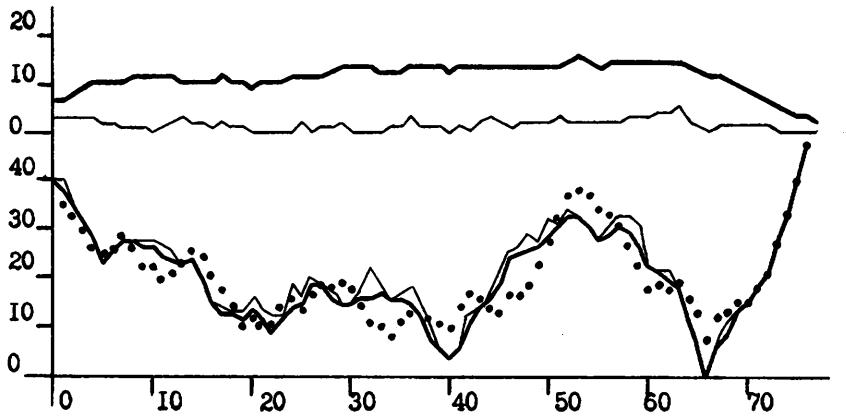

Рис. 9

Затем в качестве текста M был взят тот же комплементарный код, но в обратном порядке - ЦГАААУЦ. Сплошные кривые на рис. 9 изображают зависимости Re_ρ (внизу) и Im_ρ (вверху) от смещения для этого варианта слова M. Тонкая кривая соответствует минимальному окну, жирная - окну, расширенному на 8 позиций в обе стороны. При расширении окна количество псевдосвязей в контакте может только возрасти, а сумма длин связей, напротив, только уменьшиться. По этой причине на рис.9 жирная кривая для Im_ρ лежит выше тонкой, а для Re_ρ, наоборот, ниже. В положении $\Delta = 66$ оба варианта имеют нулевую действительную часть (в случае малого окна вообще $\rho(M,N) = 0$). Этого и следовало ожидать, поскольку реальная молекула имеет пространственную структуру "клеверного листа", где начальный участок ГАУУУЦГ соединен водородными связями в обратной последовательности с участком ЦГАААУЦ, начинающимся с 67-й позиции [3].

ПРИМЕР 6. Дискретизация плоского изображения сеткой растра превращает его в двумерный текст (см. [I]), где каждый элемент разложения - пиксель - становится вхождением, принимая значение из конечного алфавита (например - градаций яркости). Расстояние между двумерными текстами можно вводить различными способами, но экспоненциальный рост времени вычисления неприемлем для практическо-

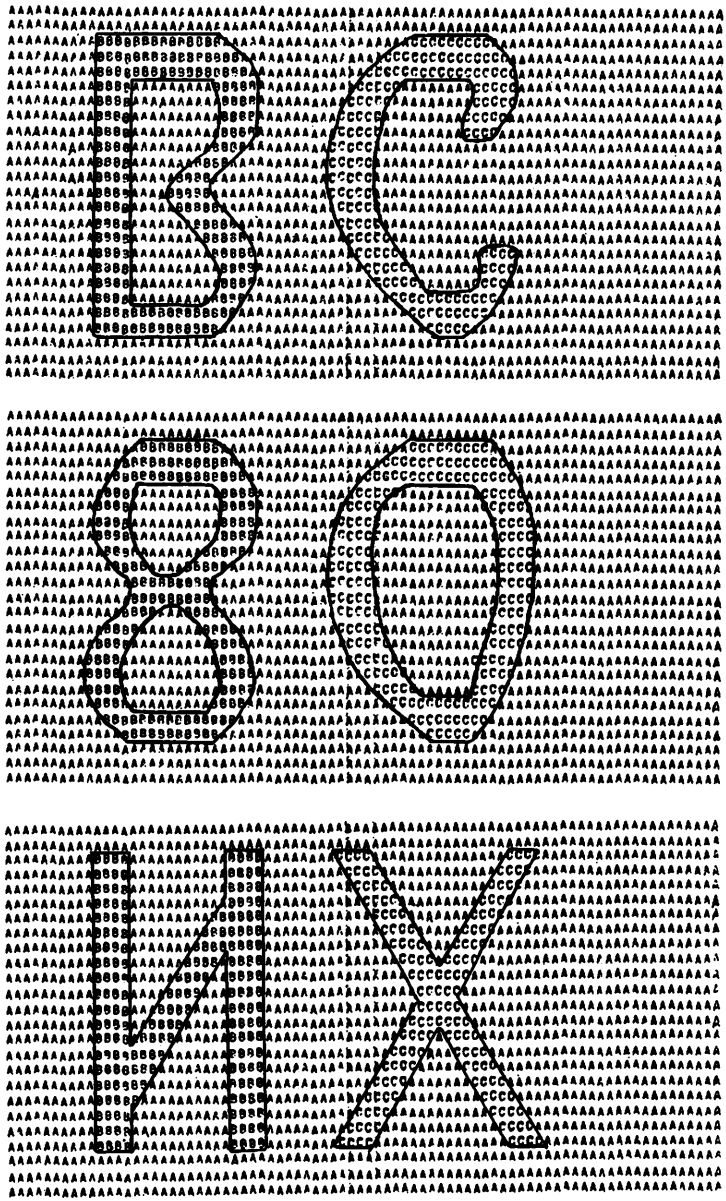

Рис. 10. Примеры текстов изображений.

го применения способа. Доступным в этом смысле является способ расщепления. Он сводит задачу вычисления расстояния между двумерными текстами к вычислению множества расстояний между одномерными текстами. На рис. 10 приведены примеры текстов изображений, для которых производилось расщепление по строкам и по столбцам. Между соответствующими строками, рассматриваемыми как одномерные тексты, вычислялись расстояния по вышерассмотренной методике. Затем то же производилось со столбцами, и все вычисленные "одномерные" расстояния суммировались. В табл.4 сведены результаты. Как и следовало ожидать, минимальное расстояние получилось в паре "ВС,80", а время вычисления на ЭВМ ЕС-1060 одного расстояния между двумерными словами размерами 25x80 пикселей - примерно 10 с.

Таблица 4

	80	ИХ	
ВС	349 + 96 i	1612 + 224i	(x)
	46 + 202 i	1180 + 326i	(y)
	395 + 298 i	2792 + 550i	(ρ)
80	0	2032 + 1741	(x)
	0	1148 + 448i	(y)
		3180 + 622i	(ρ)

Заключение

Близость текстов определяется как минимум функционала, заданного на множестве контактов текстов. Мощность множества контактов экспоненциально растет при возрастании размеров сравниваемых текстов. В статье предложен компромиссный подход к вычислению критерия близости, использующий тривиальную метрику на алфавите и метрику числовой оси для сравнения индексов вхождений букв. Это позволяет расщеплять тексты на подтексты в унарном алфавите и применять для последних методы направленного поиска минимума выпуклого функционала. Таким образом достигается полиномиальная зависимость трудоемкости вычислений от размеров сравниваемых текстов.

Литература

1. ЛЕУС В.А. Формальные критерии близости слов //Проблемы управления и теории информации. -1979.-Т.8, №4. -С. 313-325.

2. ЕГАНОВ Э.А., СОВЕТОВ Ю.К. Каратау - модель региона фосфоритонакопления. - Новосибирск: Наука, 1979. - 192 с.

3. ВОЛЬКЕНШТЕЙН М.В. Физика и биология. -М.: Наука, 1980. - 151 с.

Поступила в ред.-изд.отд.
23 октября 1987 года