ПРОБЛЕМЫ ПЕРЕДАЧИ ИНФОРМАЦИИ

Том 40 2004 Вып. 2

УДК 621.391.1 : 519.27

© 2004 г. С. А. Некрасов

ТЕОРЕТИКО-ВЕРОЯТНОСТНОЕ И СТАТИСТИЧЕСКОЕ ИССЛЕДОВАНИЕ СЛОВАРНЫХ ТЕКСТОВ

Рассматриваются теоретико-вероятностные и статистические модели и методы для расчета характеристик словарных структур. Приводятся результаты статистического исследования ряда употребительных словарей с целью проверки адекватности предложенных методов расчета.

§ 1. Введение

В задачах математического моделирования текстовых структур, при расчете различных систем обработки текстов, а также во многих других случаях может потребоваться знание соответствующих вероятностных и статистических характеристик [1–4]. Такие характеристики могут быть получены как методом имитационного эксперимента, так и с помощью теоретико-вероятностного моделирования. В качестве критериев адекватности моделей могут применяться различные статистики.

В статье рассматривается задача моделирования словарной структуры, представляющей последовательность вида:

$$\begin{array}{ll} \langle \text{ЗС}(1)\rangle & \langle \text{Статья}(1)\rangle \\ \langle \text{ЗС}(2)\rangle & \langle \text{Статья}(2)\rangle \\ \multicolumn{2}{c}{\dots\dots\dots\dots\dots} \\ \langle \text{ЗС}(N)\rangle & \langle \text{Статья}(N)\rangle, \end{array}$$

где $\langle \text{ЗС}(k)\rangle$ – k-е по порядку при счете с начала словаря заглавное слово (ЗС) входного языка, $\langle \text{Статья}(k)\rangle$ – соответствующая словарная статья (или словарное гнездо), состоящая из конечной последовательности терминальных символов входного и выходного языков.

Детерминированное прогнозирование размеров отдельных записей в реальных словарях практически невозможно. По этой причине основным эффективным способом расчета является теоретико-вероятностное и статистическое моделирование. Основным результатом исследования является вероятностный закон распределения вектора параметров, характеризующих расположение записей.

Некоторые аналогичные задачи решаются или сформулированы в теории программирования и баз данных [1, 2]. Например, в классической монографии [1] для решения ряда задач анализа машинных алгоритмов и информационных структур (например, алгоритма динамического распределения памяти) используется как строгий теоретико-вероятностный подход, так и моделирование на основе метода Монте-Карло. Однако для случая списковых структур (соответствующих рассматриваемому в статье случаю словарных текстов) в [1] исследован только простейший случай записей фиксированной длины.

В теории проектирования баз данных одной из основных задач является прогнозирование затрат памяти при хранении записей переменной длины в древовидных структурах (или составных списках) [2].

В данной статье развиваются результаты, изложенные в [3]. В частности, для обоснования адекватности предложенной модели и метода решения использованы статистические данные для более широкого множества словарей, исследовано влияние способа форматирования текста.

§ 2. Имитационное и теоретико-вероятностное моделирование

2.1. Параметризация расположения слов в тексте словаря. Будем предполагать, что текст словаря форматирован традиционным образом, т.е. постранично со стандартной сигнатурой. Расположение ЗС в тексте описывается вектором параметров, координаты которого включают значения сигнатуры, а также другие числовые характеристики.

Координаты ЗС могут быть подразделены на глобальные и локальные. К глобальным координатам относятся, в частности, порядковый номер ЗС, номер строки расположения ЗС при счете с начала словаря. Номер строки (при счете с начала словаря), в которой расположено ЗС(i), обозначим через ξ_i, а количество строк в соответствующей словарной статье – через l_i, тогда $\xi_i = x_0 + l_1 + \ldots + l_{i-1}$, $i = 2, \ldots, N(\text{ЗС})$, $\xi_1 = x_0$, где $N(\text{ЗС})$ – количество ЗС в словаре, $x_0 = \text{const}$.

Локальные координаты определяются для страницы (колонки текста), на которой расположено рассматриваемое ЗС, например, x_1 – номер строки, x_2 – порядковый номер ЗС при счете с начала соответствующей страницы (колонки текста).

Для конкретности (без ограничения общности методики расчета) будем предполагать, что ЗС(1) находится в первой строке ($\xi_1 = 1$). Относительно редкие и небольшие (в несколько строк) разрывы текста между списками слов на разные буквы будем считать фиктивными словарными статьями с соответствующей длиной.

2.2. Асимптотический закон распределения номера строки расположения ЗС. Данная характеристика относится к одной из наиболее важных в модели, так как она используется при выводе ряда статистик (см. § 3 и работу [3]). Отыскание данной характеристики практически возможно только на основе некоторых эмпирически обоснованных допущений.

На основании статистического и теоретико-лингвистического исследования словарей различных типов можно исходить из предположения, что величины l_i являются взаимно независимыми случайными величинами, распределенными по одному закону с математическим ожиданием $\mathbf{E}[l_i] = a$ и дисперсией $\mathbf{D}[l_i] = \sigma^2$. Основанием для такого предположения являются результаты проведенного автором статистического исследования множества словарей, согласно которому коэффициент корреляции длин рядом расположенных словарных статей имеет значения порядка 0,01. Для удаленных друг от друга ЗС свойство независимости длин соответствующих словарных статей является очевидным по причине различия смыслового значения, употребительности и других характеристик ЗС. В условиях сделанного предположения в силу центральной предельной теоремы [5–7] случайные величины $\xi_i, i = 1, \ldots, N(\text{ЗС})$, определенные в п. 2.1, подчиняются асимптотически нормальному закону распределения с параметрами

$$\mathbf{E}[\xi_i] = x_0 + (i-1)a, \quad \mathbf{D}[\xi_i] = (i-1)\sigma^2, \quad i = 1, \ldots, N(\text{ЗС}).$$

Рассмотрим отношение $t_i = p_i(z_1)/p_i(z_2)$, где $p_i(z)$ – плотность асимптотического распределения ξ_i при $i \to \infty$, $|z_j - m_i| \le c\sigma_i$, $j = 1, 2$, $m_i = \mathbf{E}[\xi_i]$, $\sigma_i = \mathbf{D}^{1/2}[\xi_i]$, а c – некоторая константа, выбираемая из условия, чтобы вероятность попадания значений случайной величины ξ_i в область $|z - \mathbf{E}[\xi_i]| \leq c\sigma_i$ практически равнялась нулю, вследствие чего указанную область можно не учитывать. Пусть $|z_1 - z_2| \leq$ $\leq K_{\text{line}}$, где K_{line} – количество строк на странице текста. Так как

$$p_i(z) = (2\pi)^{-1/2}\sigma_i^{-1}\exp[-(z - m_i)^2/(2\sigma_i^2)],$$

то

$$t_i = \exp[(z_2 - z_1)(z_2 + z_1 - 2m_i)/(2\sigma_i^2)] = \exp[O(1/i^{1/2})] \to 1 \quad \text{при} \quad i \to \infty,$$

поскольку

$$|(z_2 - z_1)(z_2 + z_1 - 2m_i)/(2\sigma_i^2)| \leq K_{\text{line}}c/\sigma_i = O(1/i^{1/2}) \to 0 \quad \text{при} \quad i \to \infty.$$

Таким образом, если дисперсии длин словарных гнезд достаточно велики, а страниц в словаре относительно много, то в пределах страницы или колонки текста плотность распределения величин ξ_i при $i \to \infty$ практически постоянна. По указанной причине для каждого ЗС события случайного попадания в ту или иную строку колонки текста можно считать приближенно равновероятными (в типичном случае двух колонок на одну страницу соответствующая вероятность равна $P(x_1 = j) \sim 2/K_{\text{line}}$, $j = 1, \ldots, K_{\text{line}}/2$).

Если несколько ЗС достаточно далеко отстоят друг от друга, то аналогичный вывод справедлив также для условных распределений и с высокой точностью выполняется свойство независимости их координат.

Следовательно, вероятности случайного попадания некоторого множества слов в строку с заданным номером x_1 подчиняются закону распределения, асимптотически близкому к биномиальному. Данный вывод полностью согласуется с результатами имитационного моделирования на ЭВМ с помощью метода Монте-Карло.

В программе моделирования осуществляется расчет вероятности попадания k ЗС из группы объемом n ЗС в строку с заданным номером x_1. Порядковый номер i заглавного слова (ЗС) меняется с шагом $h = 100 - 1000$. Номер строки расположения ЗС при счете с начала словаря вычислялся в соответствии с определением величины ξ_i (см. п. 2.1) и по рекуррентной формуле $\xi_{jh+1} = \xi_{(j-1)h+1} + w_j(h)$, $j = 1, 2, \ldots$, где $w_j(h)$ – независимые нормально распределенные случайные величины с математическим ожиданием ha и дисперсией $h\sigma^2$.

Статистическое исследование словарей различных типов показывает, что дисперсия длин словарных статей σ^2 является величиной порядка нескольких единиц. При таких значениях числовых параметров имеет место совпадение теоретико-вероятностных и полученных на ЭВМ экспериментальных оценок вероятностей как минимум по порядку величины.

2.3. Закон распределения номера ЗС. Для того чтобы ЗС(i) имело локальную координату $x_2 \geq j$ (событие $x_2[\text{ЗС}(i)] \geq j$) при $j \geq 2$, необходимо и достаточно, чтобы ЗС$(i - j + 1)$ находилось на той же странице, т.е. в формальных обозначениях:

$$\begin{aligned}&(x_2[\text{ЗС}(i)] \geq j) = (\text{ЗС}(i) \in \text{ Стр}(1)) \text{ И } (\text{ЗС}(i - j + 1) \in \text{ Стр}(1)) \text{ ИЛИ}\\ &(\text{ЗС}(i) \in \text{ Стр}(2)) \text{ И } (\text{ЗС}(i - j + 1) \in \text{ Стр}(2)) \text{ ИЛИ } \ldots\\ &\ldots \text{ ИЛИ } (\text{ЗС}(i) \in \text{ Стр}(K_{\max})) \text{ И } (\text{ЗС}(i - j + 1) \in \text{ Стр}(K_{\max}),\end{aligned} \tag{1}$$

где $K_{\max}$ – количество страниц в словаре, Стр(k) – страница номер k, $k = 1, \ldots, K_{\max}$.

Событие (ЗС$(i) \in$ Стр(k)) И (ЗС$(i - j + 1) \in$ Стр(k)) можно записать в равносильной форме: $(\xi_{i-j+1} > \alpha_{k-1}, \quad \xi_i \leq \alpha_k)$, где $\alpha_k = K_{\text{line}}k$. Так как $\xi_i = \xi_{i-j+1} + \eta$, где $\eta = l_{i-j+1} + \ldots + l_{i-1}$, то

$$(\text{ЗС}(i) \in \text{ Стр}(k)) \text{ И } (\text{ЗС}(i-j+1) \in \text{ Стр }(k)) = \\ = (\xi_{i-j+1} > \alpha_{k-1}, \xi_{i-j+1} + \eta \leq \alpha_k, \eta \geq 1). \tag{2}$$

Поскольку случайные величины ξ_{i-j+1} и η являются суммами независимых случайных величин, то их математические ожидания и дисперсии удовлетворяют соотношениям:

$$a_x = \mathbf{E}\xi_{i-j+1} = 1 + (i-j)a, \quad \sigma_x^2 = \mathbf{D}\xi_{i-j+1} = (i-j)\sigma^2,$$
$$a_y = \mathbf{E}\eta = (j-1)a, \quad \sigma_y^2 = \mathbf{D}\eta = (j-1)\sigma^2.$$

Неравенства в (2) выполняются с вероятностью

$$p_{j,k} = \int\limits_{\alpha_{k-1}+1}^{\alpha_k - 1} \int\limits_{1}^{\alpha_k - x} \rho_\xi(x)\rho_\eta(y)dxdy, \tag{3}$$

где $\rho_\xi(x)$ и $\rho_\eta(y)$ – плотности распределения ξ_{i-j+1} и η соответственно.

В силу (1), (2)

$$P(x_2[\text{ЗС}(i)] \geq j) = P(\xi_{i-j+1} > \alpha_0, \xi_{i-j+1} + \eta \leq \alpha_1, \eta \geq 1) + \\ + P(\xi_{i-j+1} > \alpha_1, \xi_{i-j+1} + \eta \leq \alpha_2, \eta \geq 1) + \ldots \\ \ldots + P(\xi_{i-j+1} > \alpha_{K_{\max}-1}, \xi_{i-j+1} + \eta \leq \alpha_{K_{\max}}, \eta \geq 1) = \\ = p_{j,1} + p_{j,2} + \ldots + p_{j,K_{\max}}.$$

Так как событие $x_2[\text{ЗС}(i)] \geq j$ есть сумма несовместных событий $x_2[\text{ЗС}(i)] = j$ и $x_2[\text{ЗС}(i)] \geq j+1$, то

$$P(x_2[\text{ЗС}(i)] \geq j) = P(x_2[\text{ЗС}(i)] = j) + P(x_2[\text{ЗС}(i)] \geq j+1),$$

откуда

$$P(x_2[\text{ЗС}(i)] = j) = P(x_2[\text{ЗС}(i)] \geq j) - P(x_2[\text{ЗС}(i)] \geq j+1).$$

С учетом выражения для $P(x_2[\text{ЗС}(i)] \geq j)$, полученного для произвольного $j \geq 2$, находим

$$P(x_2[\text{ЗС}(i)] = j) = \sum_{1}^{K_{\max}} (p_{j,k} - p_{j+1,k}), \quad j \geq 2. \tag{4}$$

В случае $j = 1$ расчет осуществляется аналогичным образом, но с учетом условия ЗС$(i) \in$ Стр(k) И ЗС$(i-1) \in$ Стр$(k-1)$.

Поскольку случайные величины ξ_i и η имеют аддитивную структуру и выполняются условия центральной предельной теоремы [5, с. 566; 6, с. 298], то их распределения являются приближенно нормальными:

$$\rho_\xi(x) \sim (2\pi)^{-1/2}\sigma_x^{-1} \exp[-(x-a_x)^2/(2\sigma_x^2)],$$
$$\rho_\eta(y) \sim (2\pi)^{-1/2}\sigma^{-1} \exp[-(y-a_y)^2/(2\sigma_y^2)].$$

Расчеты и асимптотический анализ соотношений (3), (4) показывают, что $P(x_2[\text{ЗС}(i)] = j) \sim a/K_{\text{line}} = \text{const}$, если j не слишком велико (достаточно выполнения условия $\rho_\eta(y) \sim 0$ при $y \geq K_{\text{line}}$. Начиная со значений $i \sim 100$, зависимостью значения вероятности $P(x_2[\text{ЗС}(i)] = j)$ от номера i практически можно пренебречь.

§ 3. Методы и результаты статистического исследования

Для проверки адекватности предложенной теоретико-вероятностной модели на основе метода статистических оценок использовались данные множества широко употребительных словарей, в частности [8–11].

В целях рандомизации исследуемые страницы словарей, словарные статьи и ЗС выбирались равновероятным случайным образом.

Результаты исследования словаря [8] следующие. Исследовались 200 из 1000 страниц текста. Контроль репрезентативности используемых статистических данных осуществлялся по критерию близости табличных оценок вероятностей и других параметров для выборок различных объемов (50, 100, 200). Установлено наличие хорошей повторяемости статистических оценок: для указанных объемов выборок расхождение в результатах составило всего несколько процентов. Эти данные также подтверждают предположения о том, что элементы выборки принадлежат одной генеральной совокупности, а случайная последовательность l_i является приближенно стационарной, вследствие чего требуемые статистические оценки с достаточно высокой точностью могут быть получены по данным одного словаря.

Обозначим через n_k количество ЗС на k-й выбранной странице, $k = 1, \ldots, N_B$; N_B – объем выборки. Из результатов анализа в п. 2.2 следует, что расположение слова в той или иной строке в пределах страницы практически равновероятно. По этой причине порядковый номер ЗС в пределах страницы или колонки текста также равномерно распределен в интервале от 1 до значения, равного количеству ЗС на данной странице или в колонке. Обозначим через μ случайную величину, равную количеству ЗС на странице, тогда с учетом вышесказанного при $i \geq I_0 = 100$:

$$P(x_2[\text{ЗС}(i)] = j) \sim \mathbf{E}[\theta(\mu - j)/\mu],$$

где $j \geq 1$, $\theta(x) = 1$, если $x \geq 0$, иначе $\theta(x) = 0$. Используя выборочные значения μ, равные $n_k, k = 1, \ldots, N_B$, можно оценить рассматриваемые вероятность и математическое ожидание через соответствующее выборочное среднее:

$$P(x_2[\text{ЗС}(i)] = j) \sim \mathbf{E}[\theta(\mu - j)/\mu] = (1/N_B) \sum_{k=1}^{N_B} \theta(n_k - j)/n_k. \tag{5}$$

Статистика (5) и соотношение (4) функционально независимы (в последнем при расчетах используются только два эмпирических момента распределения), поэтому (5) может применяться для проверки адекватности теоретико-вероятностной модели.

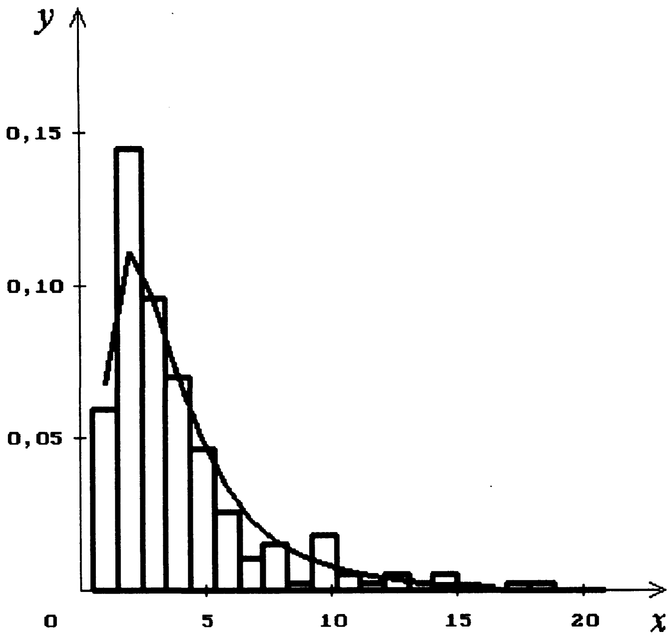
Теоретические и статистические оценки вероятностей совпадения внутристраничного номера ЗС (x_2) с различными значениями параметра j даны в таблице. Значения j, большие 30, практически маловероятны. По данным таблицы осуществлена проверка статистической гипотезы о совпадении истинных вероятностей с гипотетическими, для чего использовался соответствующий статистический критерий [7]

$$U_H = (P_c - P_m)/[P_m(1 - P_m)/N_B]^{1/2},$$

где P_c – статистическая оценка неизвестной вероятности, вычисляемая по формуле (5), P_m – теоретико-вероятностная оценка, вычисляемая в соответствии с (3), (4).

Оценки вероятности совпадений с параметром j для словаря [8]

j	$1-10$	$11-17$	$18-20$	21	$22-23$	$24-25$	26	$27-30$	Σ
P_m	0,04	0,04	0,04	0,03	0,03	0,02	0,02	0,01	1
P_c	0,05	0,04	0,03	0,03	0,02	0,02	0,01	0,01	1



Критическая точка u_k находится по таблице функции Лапласа из равенства $\Phi(u_k) = (1 - \alpha)/2$, где α – уровень значимости. Если $|U_H| < u_k$, то гипотеза равенства вероятностей может быть принята, в противном случае гипотезу отвергают.

В результате установлено, что при $j \geq 1$ и $\alpha < 0{,}6$, в том числе при стандартном значении уровня значимости $\alpha = 0{,}05$, гипотеза равенства неизвестных и гипотетических вероятностей может быть принята.

Результаты исследования словарей [9–11] в основном аналогичны. Обрабатывались выборки объемом $n = 50 - 500$.

Проведено исследование вероятностного закона распределения длин словарных статей (и, соответственно, количества информации в них). Для словаря [8] по данным выборки из 1190 элементов установлено, что по критерию Пирсона с уровнем значимости $\alpha = 0{,}05$ может быть принята гипотеза логнормальности распределения длины словарной статьи. Данная гипотеза подтверждается с высокими уровнями значимости и для словарей [9–11], что свидетельствует о закономерном характере этого явления. Графики гистограммы для выборочного и полигона для гипотетического логнормального законов распределения представлены на рисунке.

§ 4. Моделирование с учетом способа форматирования словарных текстов

Расчеты с помощью имитационных программ и обоснованной в §§ 2, 3 модели, а также данные [3] показывают, что по интегральным вероятностно-статистическим характеристикам (вероятностям совпадения с заданными значениями параметра j и т.п.) имеет место хорошее соответствие. Однако расчетные и реальные частоты количества разрывов словарных статей при переносе с одной страницы на другую для широкого множества словарей существенно разнятся. Для словаря [11] расчетная частота попадания ЗС в левый верхний угол страницы оказалась примерно равной 0,4, в то время как в действительности частота заполнения углов приблизительно равна 0,66 (погрешность около 40 процентов). Для случая словаря [8] соответствующие значения составляют 0,24 и 0,48 (расхождение 50 процентов).

Причина обнаруженного явления связана с технологическими особенностями форматирования словарных текстов: в целях удобства работы пользователя и по другим причинам создатели словарей стремятся избегать разрывов словарных статей, организуя текст так, чтобы каждая страница начиналась с новой словарной статьи. Это стремление приводит также к незначительному растяжению словарного текста и увеличению средней длины словарных статей.

Для учета данной особенности технологии форматирования словарей методика расчета была усовершенствована. Программа имитации словарного текста была дополнена условием, согласно которому ЗС, попадающее в последнюю строку на странице, переносится в начало новой страницы, если длина соответствующей словарной статьи больше одной строки. В этом случае частоты попадания ЗС в начало страниц оказались равными 0,69 для словаря [11] и 0,45 для словаря [8], т.е. погрешность составила всего 5–7 процентов, что соответствует точности статистического моделирования.

В следующей серии вычислительных экспериментов исследовалось влияние рассматриваемого технологического фактора на вероятность попадания ЗС в различные строки страницы. Было установлено, что степень влияния данного фактора в целом пренебрежимо мала. Во всех случаях наблюдалось равновероятное распределение ЗС по строкам страницы, за исключением первой и последней строк. Это объясняет факт удовлетворительного совпадения расчетных и статистических характеристик, описанный в § 3 статьи и в работе [3].

Для случая, когда осуществляется форматирование не страницы в целом, а каждой колонки текста, были получены аналогичные результаты.

§ 5. Заключение

Предложенная методика расчета вероятностей основана на использовании центральной предельной теоремы, корректность применения которой обоснована. При малых значениях параметра j суммы случайных величин содержат относительно небольшое число слагаемых, поэтому точность гауссовской аппроксимации может быть не очень высокой. Для решения данной проблемы существует подход, основанный на применении точных аналитических выражений для закона распределения суммы случайных величин. Однако недостатком данного подхода является необходимость предварительного статистического исследования закона распределения слагаемых. Этот закон, естественно, определяется по эмпирическим данным с некоторой трудно контролируемой погрешностью, которая непредсказуемым образом может сказаться на точности окончательного результата (значениях суммарных вероятностей). Рассмотренный в статье пример статистического исследования конкретных словарей показал, что точность гауссовской асимптотики является практически приемлемой. В пользу этого говорят также результаты непосредственной проверки статистических гипотез о нормальности распределения сумм моделируемых случайных величин.

В результате исследований обнаружен статистико-информационный феномен единообразия (логнормальности) закона распределения объема информации в словарных статьях.

Разработанная методика и результаты исследования могут применяться в задачах моделирования текстов на традиционных видах носителей, а также при проектировании компьютерных баз данных и редакторов текста, так как в них широко используются сегментирование записей и различные способы их форматирования [2].

СПИСОК ЛИТЕРАТУРЫ

1. *Кнут Д.* Искусство программирования на ЭВМ: Основные алгоритмы. М.: Мир, 1976.
2. *Хаббард Дж.* Автоматизированное проектирование баз данных. М.: Мир, 1984.
3. *Некрасов С.А.* Теоретико-вероятностные и статистические модели для расчета характеристик информационных систем // Изв. вузов. Электромеханика. 2003. № 4. С. 23–32.
4. *Некрасов С.А.* Феномены пространственно-смысловых корреляций в текстах // Научная мысль Кавказа. Приложение. 2003. № 11. С. 174–180.
5. *Корн Г., Корн Т.* Справочник по математике. М.: Наука, 1978.
6. *Феллер В.* Введение в теорию вероятностей. Т. 1, 2. М.: Мир, 1984.
7. *Гмурман В.Е.* Руководство к решению задач по теории вероятностей и математической статистике. М.: Высш. шк., 1999.
8. *Никонова О.Н.* Русско-немецкий словарь. Изд. 5-е. М.: Сов. энциклопедия, 1971.
9. Немецко-русский и русско-немецкий словарь для школьников. Киев: Аконит, 1995.
10. Англо-русский и русско-английский словарь для школьников и студентов. Смоленск: Русич, 1999.
11. Англо-русский и русско-английский словарь на 18 тыс. слов. М.: Рус. яз., 1968.

Некрасов Сергей Александрович
Южно-Российский государственный
технический университет, Новочеркасск
`nekrasoff_novoch@mail.ru`

Поступила в редакцию
24.07.2003

После переработки
30.01.2004