Проблемы построения алфавитного классификатора (на примере массива СУБД НИКА)

А.В. Соловьев, В.А. Тищенко

Аннотация. Рассматриваются проблемы, возникающие при построении алфавитного классификатора достаточно больших массивов текстовых ключей. По причине неравномерности распределения слов (текстовых ключей) по буквенным сочетаниям возникает проблема, связанная с построением оптимальной структуры алфавитного классификатора для перехода на заданный ключ. Рассматриваются характеристики классификатора, как случайное распределение длины ключа и распределение числа вершин в группе. Предлагается модель регрессионной зависимости с использованием ортогональных полиномов средней длины ключа в группе от максимального числа в ней вершин. Приводится пример построения такой зависимости для поля ФИО. На разных примерах зависимостей анализируется их вид и диапазон применения. Приводится пример зависимости, построенной на основе модели нечеткого регрессионного анализа.

Ключевые слова: *многоуровневый алфавитный классификатор, регрессионная зависимость, длина ключа в классификаторе, число вершин в группе.*

Введение

Использование алфавитных классификаторов является известным способом организации быстрого перехода на искомое понятие. Такой переход является альтернативой к поиску данных в ключевом текстовом массиве через поле ввода и автозаполнение [1,2]. Автозаполнение, основанное на том же подходе, что и алфавитный классификатор, было описано в статье [2] и представляет собой некоторую «динамическую» форму классификатора. При большом объеме вершин в ключевом массиве возникает проблема, связанная с построением оптимального с точки зрения числа переходов по уровням многоуровневого алфавитного классификатора. Эта проблема связана с построением регрессионной зависимости между средней длиной ключа в алфавитном указателе и средним числом вершин на этот ключ. Основу алфавитного классификатора составляет, так называемое, префиксное дерево сочетаний (ПДС), которое было рассмотрено Кнутом в его монографии «Искусство программирования» [3]. Узлами такого дерева являются буквы или сочетания букв. Каждая из вершин ключевого массива представлена в таком дереве в виде полного пути от корня до листа. Число листьев совпадает с числом вершин на ключевом уровне. Каждый узел представлен очередной буквой или буквами из текущей вершины. Ветвление в дереве происходит в случае, когда у текущей вершины в данной позиции стоит буква, отличающаяся от буквы в той же позиции у предыдущей вершины. Процесс построения ПДС содержит следующие шаги.

Шаг 1. Переход к первой вершине ключевого массива.

Шаг 2. Берется первая буква текущей вершины. Если среди вершин ПДС первого уровня отсутствует вершина с этой буквой, то она добавляется в дерево.

Шаг 3. Берется следующая буква текущей вершины. Если среди вершин ПДС соответствующего уровня отсутствует вершина с этой буквой, то она добавляется в дерево.

Шаг 4. Если есть следующая буква, то переход к шагу 3.

Шаг 5. Переход к следующей вершине ключевого массива. Если она существует, то переход к шагу 2.

Полученное в результате построения дерево будет иметь во всех узлах по одной букве вершины. Такое дерево посредством объединения последовательно идущих узлов без ветвления в один узел дает «сжатое» дерево, в котором вершины могут содержать более одной буквы. Будем рассматривать ПДС такого вида. Каждый путь от корня в ПДС задает группу вершин, начинающихся с этого ключа. Этот путь является начальным буквенным сочетанием или ключом этой группы. В общем случае распределение вершин по буквенным сочетаниям является неравномерным. Это означает, что группы с ключами равной длины могут содержать различное число вершин ключевого уровня. ПДС задает разбиение ключевого массива на группы вершин возможно различного размера с соответствующими ключами. Такому разбиению

можно противопоставить равномерное разбиение, соответствующее равномерному распределению вершин по сочетаниям. В этом случае ключевой массив разбивается на группы равного размера. Можно рассмотреть наложение двух разбиений – неравномерного в виде ПДС на равномерное с группами равного размера n. Величина n принимается как максимальный размер группы в ПДС. Исходя из величины n, можно определить среднюю длину ключа k, необходимую для перехода к искомой вершине. Меняя величину n в заданных пределах можно построить регрессионную зависимость k от n. Предварительно будет рассмотрено разбиение на группы с помощью ПДС на примере поля ФИО, а также случайные распределения длины ключа группы и числа вершин в группе. В конце рассматривается нечеткая регрессионная модель зависимости.

Работа пользователя с многоуровневым классификатором на основе ПДС выглядит следующим образом. На каждом уровне классификатора пользователь выбирает искомое сочетание букв, добавляя эти буквы к ключу группы, и с последнего уровня классификатора переходит на группу вершин в виде традиционного представления ключевого массива, как списка с искомой вершиной. Такая организация интерактивного доступа является альтернативной к поиску с помощью поля ввода. Другими словами, этот доступ может быть организован в виде меню с гиперссылками на следующий уровень классификатора, как альтернатива к набору текста на клавиатуре. Алфавитный классификатор удобно использовать на смартфоне, равно как и на любом виде устройств для доступа к ключевому уровню массива БД. Такой многоуровневый классификатор был реализован для ключевого уровня массива СУБД НИКА [2,4]. Все нижеприведенные статистические данные взяты из базы данных “За Христа пострадавшие”*.

1. Разбиение на группы с помощью ПДС

На рис.1 приводится фрагмент ключевого массива, а на рис. 2 – соответствующий фрагмент ПДС. В квадратных скобках указаны счетчики для вершин следующего уровня и терминальных вершин в поддереве для данной вершины. Если принять, что число вершин в группе не больше некоторого заданного значения $n_g \le n$ то ПДС можно обрывать на тех вершинах, для которых выполняется это условие. Полученное поддерево исходного ПДС можно рассматривать как основу классификатора. Однако здесь возникает проблема, связанная со сглаживанием неравномерности распределений вершин по начальным m-граммам. Неравномерность этого распределения порождает случайные распределения длины ключа группы и количества вершин в группе. При равномерном распределении эти величины являются достоверными и связаны строгой зависимостью:

$$k(n)=\log_a(N/n). \qquad (1)$$

Если взять число вершин в группе n=10, то фрагмент ПДС на рис. 2 будет развернут не на полную глубину. На рис.3 показан фрагмент поддерева ПДС на рис.2. Нетерминальные вершины в ПДС, для которых выполнено условие $ng \le n=10$, стали терминальными вершинами в поддереве ПДС на рис.3. К ним относятся следующие сочетания: “Габ” (3,3), “Гавердовский” (12,2), “Гавриил (В” (10,2), “Гавриил (Г” (10,2), “Гавриил (И” (10,2), “Гавриил (К” (10,2), “Гавриила (” (10,2), “Гавриленко” (10,2), “Гаврили” (7,7). После буквенного сочетания приводится в скобках значение соответствующей двумерной случайной величины (κ,ν) длины ключа группы и числа вершин в этой группе. Нетрудно видеть, что в данном фрагменте все группы с числом вершин меньше 10 и длина ключа $3 \le k \le 12$, что говорит о неравномерности распределений этих величин. Однако несколько групп имеют одинаковое число вершин в группе, а также встречаются группы с одинаковой длиной ключа. Более того, эти группы могут находиться рядом. Тогда возникают некоторые «участки равномерности» в случае присутствия нескольких подчиненных букв в ПДС для данной m-граммы. Это хорошо видно в ПДС в случае, когда дерево раскрывается на одинаковую глубину на соседних вершинах. Во фрагменте поддерева ПДС на рис.3 – это группы с ключами на “Гавриил (“):“Гавриил (А”, “Гавриил (В”, “Гавриил (Г”, “Гавриил (З”, “Гавриил (И”, “Гавриил (К”, “Гавриил (Л”, “Гавриил (П”, “Гавриил (Я”. Для этих групп выполнено условие $n_g \le n$ и одновременно они имеют одинаковые длины ключей, отличающихся одной последней буквой. Следовательно, для данной начальной m-граммы, например, “Гавриил (“ можно говорить о «локальной» равномерности распределения слов по сочетаниям в смысле выполнения условия (5).

2. Случайное распределение длины ключа группы

В общем случае распределение случайной величины к длине ключа группы является мультимодальным, например, для индекса $ФИО^{32\ 127}$ при n=1

* БД “За Христа пострадавшие”, http://martyrs.pstbi.ru

это распределение имеет 4 моды в точках 10, 5, 17, 29 в порядке убывания частотных вероятностей. Первые 3 моды соответствуют в среднем имени, фамилии и отчеству. Четвертую моду можно объяснить тем, что могут быть указаны два имени для постриженных в монашество. Максимальная мода соответствует именам. Больше всего 10-буквенных ключей, содержащих фамилию и имя, идентифицирующих любую вершину из исходного набора. Индекс ФИО$^{34\,657}$ дает такой же результат с такими же частотными вероятностями, только мода в точке 17 перемещается в точку 16. Случайное распределение числа уровней в ПДС, на которые разбивается ключ, также может служить характеристикой классификатора на основе данного ПДС. Длину ключа можно измерять не только в числе символов, но и в числе уровней в ПДС. Некоторые вершины в ПДС могут содержать более одной буквы (см.рис.2). Это связано с неравномерностью распределения слов по сочетаниям. Вершины ПДС, содержащие сочетания букв, можно условно обозначить новыми буквами α_1, α_2,… ,α_c. Полученное ПДС (назовем его обобщенным) имеет в каждой вершине по одной букве и в этом смысле ближе к равномерному случаю, т.к. будут отсутствовать вершины, которым подчинена только одна вершина, как в случае «несжатого» исходного ПДС. Длина ключа в обобщенном ПДС имеет сглаженное распределение в сравнении с распределением в исходном ПДС. В рассматриваемых примерах индексов при n=1 это распределение унимодально с модой в точке 7. Здесь длина ключа измеряется в расширенном числе символов или уровнях ПДС. В исходном ПДС длина ключа может быть больше, т.к. одна буква в обобщенном ПДС может обозначать сочетание букв в исходном дереве. При n=10 рассматриваемое условное распределение длины ключа имеет 5 мод в точках 4, 9, 17, 20, 22 в порядке убывания частотных вероятностей для обеих версий индекса по полю ФИО. В случае обобщенного ПДС условное распределение длины ключа становится бимодальным с модами в точках 4 и 7 в порядке убывания частотных вероятностей.

Рассмотрим случайное распределение длины ключа, фиксируя размер группы n, как непрерывное, исходя из следующего рассуждения Крамера [5,с.192-193]. «В приложениях к статистике величины непрерывного типа встречаются тогда, когда мы имеем дело с измерением величин, могущих принимать любое значение в некоторых пределах, например, цена товара, рост человека, размер урожая. В этих случаях величины рассматриваются как непрерывные, хотя, строго говоря, фактические данные всегда разрывны, так как каждое измерение выражается целым числом, кратным

Габриалович Вера Болиславовна
Габрияник Алексей Иванович
Габышев Степан Николаевич
Гаварин Николай Иванович
Гавердовский Владимир Львович
Гавердовский Николай Николаевич
Гавриил
Гавриил #1
Гавриил (Абалымов Николай Николаевич)
Гавриил (Владимиров Григорий Петрович)
Гавриил (Воеводин Григорий Дмитриевич)
Гавриил (Горбач Петр Дмитриевич)
Гавриил (Гур Гавриил Иванович)
Гавриил (Зверев Григорий Павлович)
Гавриил (Игошкин Иван Иванович)
Гавриил (Ильин /.../ Ефимович)
Гавриил (Кожухарев Гавриил Антонович)
Гавриил (Красновский Всеволод Витальевич)
Гавриил (Лихоманов Григорий Александрович)
Гавриил (Польшин Георгий Демьянович)
Гавриил (Яцик Григорий Петрович)
Гавриила (Гурцова Матрона Васильевна)
Гавриила (Чуркина Анисья Ивановна)
Гавриленко Андрей Елисеевич
Гавриленко Тимофей Дементьевич
Гаврилин Алексей Иванович

Рис.1. Фрагмент ключевого массива для поля ФИО$^{34\,657}$

наименьшей единице измерения данной величины. Так, цена выражается в денежных единицах, длина может быть выражена в сантиметрах, вес в килограммах и т.д. Таким образом, всегда, когда для теоретических целей величины этого рода рассматриваются как непрерывные, совершается некоторая математическая идеализация».

Крамер приводит формулу для разложения в ряд Эджворда [5,с.255] любой функции плотности. При этом он отмечает, что «на практике обычно не рекомендуется идти дальше третьего и четвертого моментов»:

$$f(x) = \phi(x) - \frac{\gamma_1}{3!}\phi^{(3)}(x) + \frac{\gamma_2}{4!}\phi^{(4)}(x) + \frac{10\gamma_1^2}{6!}\phi^{(6)}(x). \quad (2)$$

Здесь φ(x) обозначает нормальную функцию плотности*, а $\varphi^{(3)}(x)$, $\varphi^{(4)}(x)$, $\varphi^{(6)}(x)$ – ее производные 3, 4, 6 порядков соответственно, коэффициенты γ_1 и γ_2 – коэффициенты асимметрии и эксцесса соответственно. При этом предполагается, что исходная

* $\phi(x) = \Phi'(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

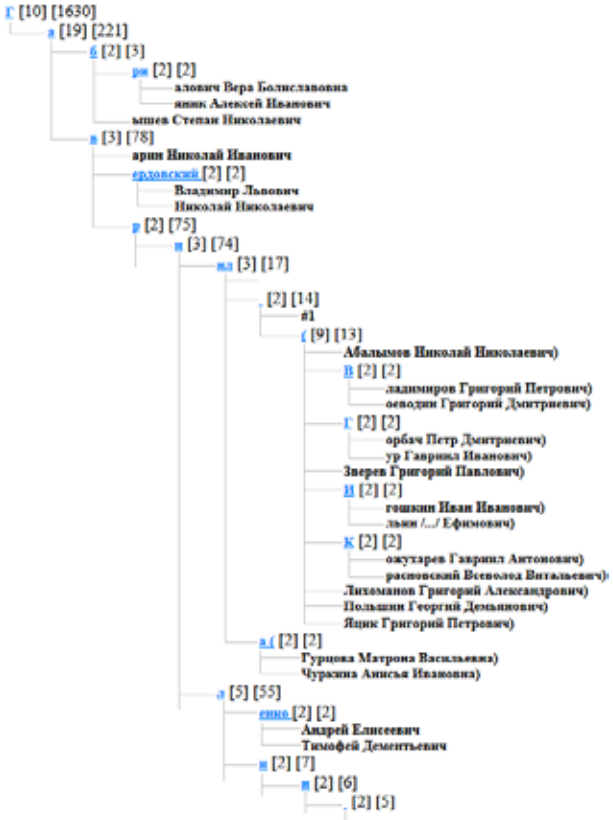

Рис.2. Фрагмент префиксного дерева сочетаний при одной вершине в группе для поля ФИО[34 657]

случайная величина нормирована. “При больших x выражение (2) иногда будет давать небольшие отрицательные значения для f(x). Это, конечно, вполне согласуется с тем, что (2) дает приближенное, но не точное выражение для функции плотности”. Кривые $\varphi^{(4)}(x)$, и $\varphi^{(6)}(x)$ симметричны относительно x=0, кривая $\varphi^{(3)}(x)$ “вводит в выражение (2) асимметричный элемент”. Формулу (2) можно применить

Рис.3. Фрагмент префиксного дерева сочетаний при 10 вершинах в группе для поля ФИО[34 657]

для аппроксимации функции плотности случайного распределения длины ключа κ, подставив выборочные моменты нормированной случайной величины длины ключа. Для поля ФИО[34 657] при одной вершине в группе выборочное среднее $m_\kappa \approx 10{,}966$ (ключ в среднем содержит 11 букв), среднеквадратическое отклонение $s_\kappa \approx 5{,}487$. В этом случае для нормированной величины $x = (\kappa - m_\kappa)/s_\kappa$ получаются следующие коэффициенты: $\gamma_1 \approx 4{,}813$ и $\gamma_2 \approx 647{,}095$, а функция плотности распределения для случайной величины x имеет вид $f(x) = \phi(x) - 0{,}802\phi^{(3)}(x) + 26{,}962\phi^{(4)}(x) + 0{,}322\phi^{(6)}(x)$.

3. Случайное распределение числа вершин в группе

Для поля ФИО[34 657] распределение числа вершин в группе близко к унимодальному с модой в точке $n_g=1$ при максимальном значении числа вершин в группе n<30, т.е. преобладают группы с одной вершиной. «Близко» означает, что в точке $n_g=1$ – глобальный максимум, а в остальных точках могут быть незначительные колебания частотных вероятностей и небольшие локальные экстремумы. При дальнейшем увеличении n случайная зависимость «сглаживается» и перестает иметь характерный максимум в точке $n_g=1$, только наблюдается тенденция уменьшения частотной вероятности с ростом n_g. При максимальном значении числа вершин в группе n=100 наблюдаются колебания частотной вероятности в произвольном виде. Дальнейшее увеличение n приводит к тому, что при значениях больших n~100 большая часть значений имеет нулевую частотную вероятность, т.е. группы с числом вершин более 100 встречаются значительно реже. По аналогии с распределением длины ключа группы можно записать функцию плотности распределения n_g, например, при максимальном значении числа вершин в группе n=1000: $f(x) = \phi(x) - 4{,}66\phi^{(3)}(x) + 113{,}673\phi^{(4)}(x) + 10{,}858\phi^{(6)}(x)$, где $x = (n_g - m_{ng})/s_{ng}$, $m_{ng}=396{,}021$ и $s_{ng}=238{,}249$.

4. Регрессионная зависимость длины ключа от числа вершин в группе

Строгая функциональная зависимость k(n) (1) для равномерного распределения слов по начальным m-граммам представлена на рис. 4.

Здесь n – число вершин в группе, $1 \le n \le N$; k – длина буквенного ключа для группы, $0 \le k \le k_m$. Любое целое значение n из области определения задает соответствующее целое значение k из области значений. В обратную сторону также верно.

В произвольном случае длина ключа *к* в классификаторе и число вершин в группе ν – случайные величины, а зависимость (1) переходит в регрессионную зависимость между матожиданиями случайных величин Мк(Мν). Для упрощения обозначения вместо букв матожиданий будем использовать буквы k и n. Исходя из соотношения (1), регрессионную модель целесообразно выбрать в виде (3) [6], заменив n на полином от n степени p:

$$k_r(n) = \ln\left(N / \left(\sum_{j=0}^{p} a_j n^j \right) \right) / \ln a . \qquad (3)$$

Такая регрессионная модель может быть сведена к общей линейной (относительно параметров) модели [7] в случае, когда степень полинома p является известной. В данном случае степень полинома p считается неизвестной и также является параметром модели. В [8] отмечается, что в такой постановке требуется оценить объект нечисловой природы $(p,a_0,a_1,a_2,\ldots,a_p)$. Обычные методы оценивания для него неприменимы, т.к. p – дискретный параметр. Подчеркивается, что существующие методы оценивания степени полинома являются практическими и не дают обязательно оптимального значения p. Одним из таких методов является метод последовательного повышения степени полинома с проверкой адекватности модели по F-критерию Фишера. Этот метод используется в данной работе для построения полиномиальной регрессии на основе формулы (3), преобразованной так, чтобы в правой части равенства стоял полином.

Рассмотрим пример данных для величин n и k по полю ФИО[34 657].

Для упрощения расчетов рассмотрим в качестве n максимальные значения вместо выборочных средних.

Как видно из табл. 1, 2 значения величины n являются целыми и увеличиваются с фиксированным шагом. В случае равноотстоящих значений удобно использовать регрессию на ортогональных полиномах [7], преобразовав выражение (3) к полиному от n:

$$y_r(n) = \sum_{j=0}^{p} a_j n^j / C = \sum_{j=0}^{p} a_{cj} n^j . \qquad (4)$$

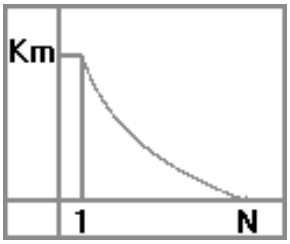

Рис.4. Зависимость k(n)

Здесь выполнено преобразование переменной k к переменной y: $e^{\ln N - \ln ak(n)} / C = y(n)$ или

$$a^{k_m - k_r(n)} / C = y_r ; \;\; k = k_m - \log_a(Cy_r). \qquad (5)$$

Здесь $k_m=\log_a N$. В рассматриваемом случае k_m=255 символов*, а число букв в алфавите принимается равным a=30. Константа C выбирается так, чтобы y менялось в пределах от сотых до нескольких тысяч. В данном случае $C=10^{368}$. Преобразование (5) приводит к полиномиальной регрессионной зависимости (4) $y_r(n)$. В табл. 3 приводятся данные для преобразованной переменной y в зависимости от x. Здесь

$$x = \frac{n}{10} - \frac{11}{2} \qquad (6)$$

является свободной переменной ортогональных полиномов, значения которой получаются из таблиц значений ортогональных полиномов [9] при 10 измерениях переменной y, вычисленной по формуле (5) при подстановке в качестве k_r значений k из табл. 2 (см.выше).

Уравнение регрессии, выраженное через ортогональные полиномы $\Psi_j(x)$, принимает вид:

$$y_r(x) = \sum_{j=0}^{p} b_j \Psi_j(x) . \qquad (7)$$

Оценки неизвестных коэффициентов определяются методом наименьших квадратов по следующим формулам:

* В СУБД НИКА максимальная длина текстового ключа равна 255 символов

Табл. 1

Данные для величин n и k по полю ФИО[34 657]

n	4,826	9,223	13,947	17,964	22,681	27,412	30,732	34,154	38,361	42,650
k	6,703	5,717	5,142	4,799	4,490	4,244	4,082	3,959	3,817	3,690

Табл. 2

Данные для величин n и k по полю ФИО[34 657]

n	10	20	30	40	50	60	70	80	90	100
k	6,703	5,717	5,142	4,799	4,490	4,244	4,082	3,959	3,817	3,690

$$b_j = \frac{1}{S_j^2}\sum_{i=1}^{m} y(x_i)\Psi_j(x_i), \text{ где } \; S_j^2 = \sum_{i=1}^{n}\Psi_j^2(x_i) \quad (8)$$

Здесь $\Psi_j(x_i)$ обозначены значения ортогональных полиномов степени j [9]. Оценки коэффициентов b_j имеют дисперсию σ^2/S_j^2. Неизвестная дисперсия σ^2 имеет эффективную оценку в виде оценки остаточной дисперсии, определяемую формулой [9]:

$$s^2(p) = \frac{1}{m-p-1}\left(\sum_{i=1}^{m} y_i - \sum_{j=1}^{p} b_j^2 S_j^2\right). \quad (9)$$

В случае неизвестной степени p аппроксимирующего полинома уравнения регрессии (7) в [10] рекомендуется определять p путем последовательных уточнений. Критерием для прекращения процедуры уточнения степени полинома является величина остаточной дисперсии (9). Если $s^2(p+1)>s^2(p)$, то в качестве регрессии принимается полином степени p. Значимость дисперсий $s^2(p+1)$ и $s^2(p)$ проверяется F-критерием Фишера при m-p и m-p-1 степенях свободы. В статье о методах снижения размерности [11] приводится состоятельная оценка величины p, как выражение от среднего квадрата ошибки α_p: $p^* = Arg\min_p\{\alpha_{p+1} - 2\alpha_p + \alpha_{p-1}\}$. Однако в [8] подчеркивается, что "излишнее усложнение статистических моделей вредно" и отмечается, что оценка остаточной дисперсии $s^2(p)$ будет колебаться около предела $\lim_{m\to\infty} s^2(p) = \sigma^2$. Поэтому в качестве оценки неизвестной степени многочлена можно использовать, например, первый локальный минимум оценки остаточной дисперсии $s^2(p)$, несмотря на то, что эта оценка степени полинома не является состоятельной и предельное распределение этой оценки является геометрическим [8]. В результате применения формул (8) для определения коэффициентов ортогональных полиномов и формул для остаточных дисперсий можно получить табл. 4.

Из предпоследней строки табл. 4 для b_6 следует, что $S_5^2/S_6^2 \approx 0.790 < 1$. Это означает, что в качестве оценки степени полинома можно взять пятую степень, т.к. она соответствует первому локальному минимуму остаточной дисперсии. С помощью F-критерия Фишера этот результат можно уточнить. Из последней строки видно, что $S_3^2/S_4^2 \approx 2{,}017 < F_{0,95}(m\text{-}p, m\text{-}p\text{-}1) = F_{0,95}(5,4) = 4{,}95$. Это означает, что остаточная дисперсия уменьшается незначимо при переходе от третьей степени к четвертой. Следовательно, аппроксимирующий полином имеет третью степень. По критерию Стьюдента при уровне значимости α=0,05 все коэффициенты регрессии являются значимыми (см. неравенство (10) и табл. 5):

$$|b_j| / S_{bj} > t(6;0{,}975) = 2{,}447. \quad (10)$$

В результате получаем регрессионную зависимость:

$$y_r(x) = 420{,}713 + 81{,}159\Psi_1(x) + 64{,}209\Psi_2(x) + 2{,}04\Psi_3(x) \quad (11)$$

В табл. 6 приводятся данные для y_r, полученные из модели (11).

Проверка статистик по остаткам [7] основывается на предположении, что остатки должны иметь сходство с наблюдениями из нормального распределения (для применения F-критерия) со средним, равным нулю. Среднее остатков равно $\sum r_i / m = 0$, также статистика $T_{11} = \sum r_i y_{ri} \approx 0$ (линейный член правильно включен в модель). В [8] отмечается, что модели на основе нормального распределения, как правило, неадекватны реальной ситуации, но позволяют глубже изучить суть рассматриваемого явления и пригодны для первоначального анализа. Следовательно, в рассматриваемом случае правильно принять не термин "адекватность модели", а термин "модель с меньшей ошибкой" в смысле критерия Фишера в предположении, что остатки распределены нормально с нулевым средним и одинаковой дисперсией. Регрессионная зависимость описывает разброс данных относительно среднего <y>=420,7 (k=4,091) на R^2*100%=99,86% (квадрат множественного коэффициента корреляции между y и y_r) . В результате проверки полезности регрессии [7] методом Бокса-Уэтса получаем $(y_{r\max} - y_{r\min})/(ps^2/n)^{1/2} \approx 4{,}048 \geq 4$ и практическое правило "четырехкратного превышения" выполняется.

После подстановки выражений для ортогональных полиномов [9] получаем зависимость: $y_r(x) = 155{,}849 + 112{,}499x + 32{,}105x^2 + 3{,}401x^3$.

При переходе (6) от переменной x к переменной n получаем зависимость $y_r(n) = 0{,}003n^3 - 0{,}240n^2 + +6{,}798n - 57{,}561x^3$.

Наконец, с помощью обратного преобразования (5) получаем регрессионную зависимость $k_r(n)$ и с помощью границ 95%-го доверительного интервала зависимости $k_{rmin}(n)$ и $k_{rmax}(n)$ (см. последнюю строку в табл. 5):

$$k_r(n) = 5{,}867 - \log_{30}(0{,}0034n^3 - 0{,}240n^2 + 6{,}798n - 57{,}561) \quad \text{при } 14 < n \leq 1000 \quad (12)$$

$$k_{r\min}(n) = 5{,}867 - \log_{30}(0{,}0029n^3 - 0{,}179n^2 + 5{,}889n - 100{,}258) \quad (12a)$$

$$k_{r\max}(n) = 5{,}867 - \log_{30}(0{,}0045n^3 - 0{,}397n^2 + 12{,}882n - 106{,}176) \quad (12b)$$

а также аналогичным методом получаем зависимость для среднего числа уровней алфавитного классификатора от максимального числа вершин в группе:

Табл. 3

Данные для переменных x и y

x	-9/2	-7/2	-5/2	-3/2	-1/2	1/2	3/2	5/2	7/2	9/2
y	0,058	1,664	11,764	37,776	108,055	249,469	432,824	657,653	1065,978	1641,884

Табл. 4

Оценки коэффициентов уравнения регрессии

j	0	1	2	3	4	5	6	7
b_j	420,713	81,159	64,209	2,040	0,886	1,013	-0,243	-0,144
S_j^2	306380,499	72969,991	5649,474	637,557	316,042	194,783	246,675	66,074
S_{j-1}^2/S_j^2			12,916	8,861	2,017	1,623	0,790	3,733
$F_{0,95}$			3,73	4,21	4,95	6,26		

Табл. 5

Значимость коэффициентов регрессии и их 95%-ые доверительные интервалы

J	0	1	2	3
b_j	420,713	81,159	64,209	2,040
b_j/S_{bj}	52,690	58,390	29,216	7,485
$b_j \pm t(6;0,975)S_{bj}$	[401,2; 440,3]	[77,8; 84,6]	[58,8; 69,6]	[1,4; 2,7]

Табл. 6

Данные по остаткам

y	0,058	1,664	11,764	37,776	108,055	249,469	432,824	657,653	1065,978	1641,884
y_r	-10,165	9,581	22,120	47,859	107,201	220,550	408,310	690,886	1088,682	1622,101
r	10,223	-7,916	-10,356	-10,083	0,854	28,919	24,513	-33,233	-22,704	19,783

$$k_r(n) = 5{,}867 - \log_{30}(0{,}422n^2 - 17{,}306n + 178{,}517) \quad (12')$$
$$\text{при } 20 < n \le 1000 .$$

Аппроксимирующий полином в (12') имеет меньшую степень, чем в (12). Это можно объяснить тем, что распределение среднего числа уровней имеет "сглаженное" распределение (меньшее число мод) в сравнении с распределением средней длины ключа.

Из формулы (12) получаем следующие данные для k_r (r_k=k-k_r).

На рис.5. представлен график регрессионной зависимости $k_r(n)$ (12) в диапазоне $15 \le n \le 1000$. При этом средняя длина ключа уменьшается от 6,1 до 1,5. Схематично показана область*, соответствующая 95%-му доверительному интервалу, внутри которой располагается регрессионная зависимость $k_r(n)$. Линия над зависимостью $k_r(n)$ соответствует $k_{rmin}(n)$ (12a), линия под зависимостью $k_r(n)$ соответствует $k_{rmax}(n)$ (12b).

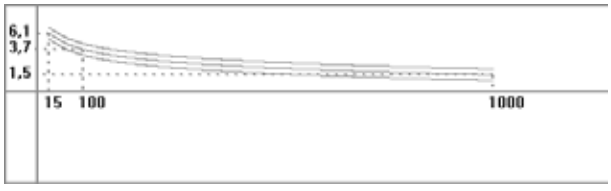

Рис.5. Регрессионная зависимость $k_r(n)$

Если округлять не до тысячных долей, а до миллионных, то можно достичь большей точности приближения. Значение k_r=5,867 при n=10 отличается от исходного на большую величину, чем остальные значения по причине того, что аппроксимирующий полином дает значение -10,165 (см. табл. 6), выпадающее из области определения обратного преобразования (7), т.к. выражение под логарифмом должно быть строго большим нуля. Значение под логарифмом в этой точке взято за единичное. Решение этой проблемы возможно с помощью использования метода наименьших квадратов при наличии ограничений [7]. В каждой точке x_i, в которой оказывается отрицательное значение полинома $y_r(x_i)$, к искомому регрессионному уравнению (9) добавляется соответствующее ограничение $y_r(x_i)>0$. Однако представля-

* Рисунок не отражает сужение доверительной области при увеличении n. Сужение происходит по причине того, что аппроксимирующий полином стоит под логарифмом.

Табл. 7

Экспериментальные и модельные значения длины ключа при различном максимальном числе вершин в группе

n	10	20	30	40	50	60	70	80	90	100
k	6,703	5,717	5,142	4,799	4,490	4,244	4,082	3,959	3,817	3,690
k_r	5,867	5,321	5,152	4,953	4,677	4,426	4,220	4,048	3,903	3,777
r_k	0,836	0,396	-0,010	-0,154	-0,187	-0,182	-0,138	-0,089	-0,086	-0,087

Табл. 8

Экспериментальные и модельные значения длины ключа вне диапазона регрессионной зависимости

n	100	200	300	400	500	600	700	800	900	1000
k	3,690	3,019	2,678	2,366	2,272	2,147	2,071	1,966	1,916	1,916
k_r	3,777	3,026	2,626	2,351	2,142	1,973	1,831	1,709	1,601	1,506
r_k	-0,087	-0,007	0,052	0,015	0,13	0,174	0,240	0,257	0,314	0,410

ется более эффективным в данном случае использовать более частный метод, который сводится к тому, чтобы рассмотреть другой диапазон и шаг, для построения регрессии. Например, можно взять 2≤n≤20 с шагом 2 и построить регрессионную зависимость (12'') отдельно для этих точек:

$$k_r(n) = 10{,}606 - \log_{30}(220{,}5625n^4 - 3288{,}5n^3 + 22127{,}12n^2 - 61711n + 55396{,}8) \quad (12'')$$
$$\text{при } 3 < n \le 100.$$

Рассмотрим значения величин n и k, отличные от тех, по которым строилась регрессионная зависимость. Точки внутри заданного диапазона 10≤n≤100 будут давать значения величин k, близкие к экспериментальным, например, n=65 дает k_r=4,318, что хорошо согласуется с экспериментальным значением k=4,163 и r_k=-0,155. Однако построенная зависимость не дает правильных значений при n<15, т.к. при этих значениях n аппроксимирующий полином, стоящий под логарифмом, принимает отрицательные значения.

Интересно рассмотреть точки вне заданного диапазона при n>100. В следующей таблице сведены данные для 100≤n≤1000 (табл. 7, 8).

Величина ошибки r_k вначале является незначительной и близка к тем значениям, которые получаются внутри диапазона. Однако с n=500 величина r_k начинает заметно расти так, что построенной регрессионной зависимостью (12) не рекомендуется пользоваться при n>1000.

Зависимости $k_r(n)$, подобные (12), можно построить для других индексируемых полей БД с текстовым ключом, например, для ФИОЗаявителя[1938] и ФИОРодства[2596*]. Используя описанный метод, основанный на ортогональных полиномах, получаются следующие зависимости. Для поля ФИО Заявителя[1938]:

$$k_r(n) = 5{,}867 - \log_{30}(34034{,}6n - 677668) \quad (13)$$
$$\text{при } 19 < n \le 300$$
$$k_r(n) = 5{,}867 - \log_{30}(33461{,}8n - 630351) \quad (13')$$
$$\text{при } 18 < n \le 300$$

и для поля ФИО Родства[2596]:

$$k_r(n) = 5{,}867 - \log_{30}(146{,}045n^2 + 1115{,}25n - 51661) \quad (14)$$
$$\text{при } 15 < n \le 300$$
$$k_r(n) = 5{,}867 - \log_{30}(147{,}055n^2 + 849{,}55n - 36691) \quad (14')$$
$$\text{при } 12 < n \le 300$$

Здесь для примера приводятся также зависимости (13') и (14') среднего числа уровней алфавитного классификатора от максимального числа вершин в группе. Эти зависимости близки к соответствующим зависимостям (13) и (14) для средней длины ключа. Эти зависимости начинают совпадать с (13') и (14') соответственно при n≥40 для обоих полей. Как видно из формул (12), (13), (14) степень полинома под логарифмом возрастает с увеличением числа вершин на ключевом уровне. Вероятно, это можно объяснить увеличением разброса случайной величины длины ключа при большем числе вершин. Соответственно и верхняя граница диапазона для случайной величины n максимального числа вершин в группе, при которой применимы регрессионные зависимости (13), (14), уменьшается с 1000 до 300. Нижняя граница величины n определяется как максимальное целое положительное значение n (меньшее верхней границы), при котором значение полинома, стоящего под логарифмом будет не больше нуля. Для получения регрессионных зависимостей при n, меньших нижней границы, можно использовать указанный выше метод построения регрессии на ортогональных полиномах на соответствующем диапазоне 2≤n≤20 с шагом 2. Из примера зависимости (12'') для этого диапазона видно, что на сте-

* В верхнем индексе указано количество элементов на ключевом уровне соответствующего массива

пень аппроксимирующего полинома может также влиять диапазон свободной переменной n и шаг изменения этой переменной. Особенностью зависимости среднего числа уровней (12') является то, что полином под логарифмом имеет экстремум в окрестности точки n=21 (20<n<22) и зависимость $k_r(n)$ в этой точке достигает максимального значения на целых значениях n. Таким образом, для того, чтобы сохранить убывающий характер зависимости $k_r(n)$, следует рассматривать $20<n\leq 1000$.

5. Уточнение регрессионной зависимости $k_r(n)$ на основе нечеткого регрессионного анализа

Полученные выше остатки [7] r в четком регрессионном анализе принимаются как ошибка наблюдения, которая является случайной величиной. В нечетком регрессионном анализе [12] остатки рассматриваются как обусловленные нечеткостью структуры модели. Среди рассмотренных в [12] нечетких моделей была взята модель нечеткого регрессионного анализа по критерию минимальной нечеткости, комбинированного с МНК, т.к. в качестве исходных данных эта модель может иметь результаты четкого регрессионного анализа, рассмотренного выше. А именно, в полученном полиноме все, коэффициенты могут быть рассмотрены как нечеткие коэффициенты с симметричной треугольной функцией принадлежности, причем в качестве нечетких центров выбираются значения коэффициентов из четкой регрессионной зависимости, а значения нечетких разбросов определяются по критерию минимальной нечеткости. В результате применения критерия для рассматриваемого выше примера зависимости для поля ФИО[34 657] получаем следующую нечеткую модель:

$$k_r(n) = (5{,}867;0) - \\ -\log_{30}((0{,}003;0)n^3 - (0{,}240;0)n^2 + (6{,}798;0)n - (57{,}561;33{,}233)) \quad (12''')$$

Зависимость (12'''), построенная на основе нечеткой регрессионной модели с нечеткими коэффициентами с симметричной треугольной функцией принадлежности, представляет собой альтернативу по отношению к доверительной области, ограниченной зависимостями $k_{rmin}(n)$ (12a) и $k_{rmax}(n)$ (12b).

6. Актуальность проблем построения алфавитного классификатора.

Решение рассмотренных проблем построения многоуровневого алфавитного классификатора позволяет построить оптимальным образом работу с ключевым массивом. Использование интерактивного интерфейса на основе такого классификатора дает возможность быстро находить искомую вершину ключевого массива без использования поля ввода. Для пользователя алфавитный классификатор представляет собой "путеводитель" или систему подсказок для перемещения к искомой вершине ключевого массива. С точки зрения базы знаний алфавитный классификатор можно рассматривать как некоторую своеобразную "схему" ключевого массива.

Литература

1. *Bast H., Weber I.* Type less, find more: fast autocompletion search with a succinct index // Proc. of SIGIR'06 conference. August 6-11, 2006. P. 364-371.
2. *Тищенко В.А.* Применение автозаполнения для перехода по ключевым словам на искомые значения в массиве СУБД НИКА // Материалы XXIII Ежегодной богословской конференции ПСТГУ. Т.1. 2013. с. 325-328.
3. *Кнут Д.Э.* Искусство программирования. Том 3. Сортировка и поиск / Пер. с англ. М.: Вильямс, 2013. Т. 3. 832 с. (Knuth D.E. The Art of Computer Programming. Sorting and Searching. – 2-nd ed. – N.Y.: Addison-Wesley, 1998. Vol.3. 782 p.).
4. *Годунов А.Н., Емельянов Н.Е., Косьмынин А.Н., Солдатов В.А.* СУБД НИКА // Системы управления базами данных и знаний. М.: Финансы и статистика. 1991. с.208-249.
5. *Крамер Г.* Математические методы статистики. / Пер. с англ. – М.: Мир, 1975. 648 с. (Cramér H. Mathematical Methods of Statistics. Princeton: Princeton University Press, 1946. 575 p.)
6. *Емельянов Н.Е., Тищенко В.А.* Методология построения многоуровневого индекса ключевого массива по лексикографическому признаку на основе метода регрессионного анализа на примере СУБД НИКА // Обработка информационных и графических ресурсов / Труды ИСА РАН. Т.58. 2010. С. 6-17.
7. *Дрейпер Н., Смит Г.* Прикладной регрессионный анализ. В 2-х кн. – М.: Финансы и статистика, 1986. (Draper N.R., Smith H. Applied regression analysis. – 2nd ed. – N.Y.: John Wiley & Sons, 1966).
8. *Орлов А.И.* Прикладная статистика. Учебник. М.: Экзамен. 2005. 672 с.
9. *Большев Л.Н., Смирнов Н.В.* Таблицы прикладной статистики. М.: Наука. 1983. С.416.
10. *Кобзарь А.И.* Прикладная математическая статистика. Для инженерных и научных работников. М.: Физматлит. 2006. С.816.

11. *Орлов А.И., Луценко Е.В.* Методы снижения размерности пространства статистических данных // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2016. № 119.

12. *Могиленко А.В.* Теория нечетких множеств. Нечеткий регрессионный анализ. Томск: Печат. Мануфактура. 2004. С.61.

Соловьев Александр Владимирович. Главный научный сотрудник ИСА ФИЦ ИУ РАН. Закончил МГТУ им. Н.Э. Баумана в 1993 г. Доктор технических наук. Количество печатных работ: более 30. Область научных интересов: фундаментальные проблемы организации электронного документооборота. E-mail: soloviev@isa.ru

Тищенко Владимир Александрович. Научный сотрудник ИСА ФИЦ ИУ РАН. Закончил МИФИ в 1993г. Количество печатных работ: 17. Область научных интересов: средства создания и поддержки электронных библиотек и электронных изданий. E-mail: vtischenko@isa.ru

The problems of constructing of alphabetical classifier (on an example of an array of NIKA DBMS)

Soloviev A.V., Tishchenko V.A.

Abstract. The problems arising in the construction of an alphabetic classifier of large enough arrays of text keys are considered. Because of the uneven distribution of words (text keys) in alphabetic combinations, there is a problem associated with constructing the optimal structure of an alphabetic classifier for switching to a given key. The characteristics of the classifier, such as the random distribution of the key length and the random distribution of the number of vertices in a group are considered. A regression dependence model of average key length in a group of the maximum number of vertices in a group using orthogonal polynomials is proposed. An example of constructing such a dependence for the field name is given. On different examples of dependencies, their type and range of applications are analyzed. An example of a dependence constructed on the basis of a model of fuzzy regression analysis is given.

Keywords: *multilevel alphabetic classifier, regression dependence, key length in the classifier, number of vertices in the group.*

References

1. Bast H., Weber I. 2006. Type less, find more: fast autocompletion search with a succinct index. The 29th annual international ACM SIGIR conference on Research and development in information retrieval. Proceedings. Seattle. 364–371.
2. Tischenko V.A. 2013. Primenenie avtozapolneniya dlya perehoda po klyuchevim slovam na iskomie znacheniya v massive SUBD NIKA [Application of autocomplete to navigate by keywords to the desired values in the NIKA database]. Materiali XXIII Ezhegodnoj bogoslovskoj konferentsii PSTGU [The XXIII Annual theological conference of the PSTGU. Proceedings]. 1:325–328.
3. Knuth D.E. 1998. The Art of Computer Programming. Sorting and Searching. 2-nd ed. N.Y.: Addison-Wesley. Vol.3. 782 p.
4. Godunov A.N., Emel'yanov N.E., Kos'minin A.N., Soldatov V.A. 1991. SUBD NIKA [NIKA system]. Sistemi upravleniya bazami dannih i znanij [Database and knowledge management systems]. M.: «Finansi i statistika». 209–248.
5. Cramér H. 1946. Mathematical Methods of Statistics. Princeton: Princeton University Press. 575 p.
6. Emelyanov N.E., Tischenko V.A. 2010. Metodologiya postroeniya mnogourovnevogo indeksa klyuchevogo massiva po leksikograficheskomu priznaku na osnove metoda regressionnogo analiza na primere SUBD NIKA [Methodology for constructing a multilevel index of a key array based on the lexicographic characteristic based on the regression analysis method on the example of the NIKA database]. Trudy ISA RAN "Obrabotka informatsionnih i graficheskih resursov" [ISA RAS "Processing of information and graphics resources" Proceedings]. 58:6–17.
7. Draper N.R., Smith H. 1966. Applied regression analysis. 2nd ed. N.Y.: John Wiley & Sons.
8. Orlov A.I. 2005. Prikladnaya statistika [Applied statistics]. M.: Exam. 672 p.
9. Bolshev L.N., Smirnov N.V. 1983. Tablitsi prikladnoj statistiki [Tables of applied statistics]. M .: Nauka. 416 p.
10. Kobzar A.I. 2006. Prikladnaya matematicheskaya statistika. Dlya inzhenernih i nauchnih rabotnikov [Applied mathematical statistics. For engineering and scientific workers]. M.: Fizmatlit. 816 p.
11. Orlov A.I., Lutsenko E.V. 2016. Metodi snizheniya razmernosti prostranstva statisticheskih dannih [Methods for reducing the dimensionality of the space of statistical data] Politematicheskij setevoj elektronnij nauchnij zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta [Polymatic Network Electronic Journal of the Kuban State Agrarian University]. N119. Available at: http://ej.kubagro.ru/2016/05/pdf/05.pdf (accessed February 2, 2018).
12. Mogilenko A.V. 2004. Teoriya nechetkih mnozhestv. Nechetkij regressionnij analiz [The theory of fuzzy sets. Fuzzy regression analysis]. Tomsk: Pechat. Manufaktura. 61 p.

Soloviev Alexander Vladimirovich. d.t.s., Main researcher, ISA FRC CSC RAS. Graduated from the MSTU N.E. Bauman in 1993. The number of printed works is more than 30. Research interests: fundamental problems of organization of electronic document management. E-mail: soloviev@isa.ru

Tishchenko Vladimir Alexandrovich. Researcher, ISA FRC CSC RAS. Graduated from the MEPhI in 1993. Number of publications: 17. Research interests: means of creation and support of electronic libraries and electronic publications. E-mail: vtischenko@isa.ru